

Ill-advised 'freedom' of scientific information

The sharing of data by researchers ought to be encouraged. But a compulsion to release raw data and notes in current US openness laws is the wrong way to achieve it, as is a proposed amendment.

The principle of granting maximum public access to government records, as embodied in the Freedom of Information Act (FOIA), is a sound one which has strengthened democracy in the United States, and from which secretive regimes elsewhere in the world have much to learn. However, a new law, passed last October, that would shed similar light on scientific records, threatens to undermine academic research, while contributing nothing to open government.

Like so many of Washington's finest ruses, the new legislation was passed in the dead of night, without hearings or outside consultation. Senator Richard Shelby (Republican, Alabama) had it quietly inserted into last October's unwieldy, 4,000-page omnibus spending act, as a prerequisite for the funding of the White House's small but powerful Office of Management and Budget (OMB). The sponsors of the measure were apparently concerned at the time about the unwillingness of researchers at the Harvard School of Public Health to release raw data behind an epidemiological study of the health effects of small carbon particles ('particulates'), which was used by the Environmental Protection Agency as a basis for regulation.

The act itself says that OMB should amend an existing circular to ensure that "all data" produced under a federal research grant be made available to the public through the procedures established under FOIA. That set off alarm bells in the scientific community, which feared that FOIA requests might be used to obtain scientific data even before publication of research findings — opening the door to the merciless harassment of scientists by hostile third parties.

Last week the OMB issued a draft amendment that would make data accessible under FOIA only after research findings had been published and used by the government "in developing policy or

rules" (see page 459). The OMB proposal still casts a wide net, however. A great deal of scientific research is arguably used by the government in that way. The sequestration and re-interpretation of a scientist's notes and computer records after publication may prove to be only marginally less disruptive than it would have been beforehand. Some scientists will avoid controversial fields of endeavour, such as pollution-related research, if the controversy is accompanied by the threat of inquisitions into their records and methods by well-financed special interest groups. Patients and commercial partners may doubt that the exemptions allowed by FOIA will protect their privacy.

The proposed change would also pull the academic research enterprise more closely under the wing of the federal government. The US Supreme Court has properly resisted the notion that academics on federal grants are agents of government, or that they be subject to the bureaucracy which enables government departments to comply with laws such as FOIA. Such subservience to government would quickly stifle the freedom on which the United States has built its scientific strength. That factor seems to have been ignored by Shelby and Trent Lott (Republican, Mississippi), the leader of the Senate, in their rush to legislate.

Universities and scientific societies are hoping that by 5 April, when its public consultation period ends, OMB will receive a loud and clear message from the scientific community about its proposed rule. That message should be that the law needs to be repealed, while acknowledging a need for an enhanced availability of primary data. The development of mechanisms to achieve that, with the full involvement of the scientific community, holds far more promise than the draconian measure which has unfortunately been passed into law. □

Planetary persistence

Debate about labels can be both powerful and pointless.

So Pluto will not be counted among the asteroids and other second-class citizens of the solar system. After weeks of e-mailed arguments among planetary scientists and media reports that mostly milked the episode for laughs, the Small Bodies Names Committee of the International Astronomical Union (IAU) decided last week not to assign Pluto a minor planet number. The smallest of the Sun's nine planets is therefore spared the indignity of what would undoubtedly have been an ill-advised demotion.

The controversy's originator, Brian Marsden, has for years unsuccessfully tried to persuade fellow astronomers that Pluto should be counted as a minor planet. As director of the Harvard Smithsonian Astrophysical Observatory's Minor Planet Center in Cambridge, Massachusetts, he proposed that the upcoming designation of the 10,000th minor planet, whose cataloguing is the centre's responsibility, be given to Pluto — partly as an honour, and also to recognize that the icy planet just as properly belongs to a class of Trans-Neptunian

Objects (TNOs) orbiting at the far reaches of the Solar System.

It was unfortunate that the proposal was erroneously linked in some press stories with the work of two IAU committees currently considering a numbering system for TNOs and the scientific definition of a planet, and even more so that some media began reporting that Pluto had already been downgraded. The community felt obliged to respond and, finally, IAU General Secretary Johannes Andersen slammed the door shut by issuing his own statement that the naming committee had squashed the suggestion.

From the beginning, most planetary astronomers never considered the matter controversial in a scientific sense, and will be happy to get back to work. Marsden appears contrite about his fruitless stirring. But his misjudgement was not scientific: the distinctions between Pluto and TNOs appear insignificant and need to be clarified. Marsden merely underestimated astronomers' proprietorial attachment to this lonely object. □



Europe bids to pull US patent law into line with first-to-file system

[LONDON] The European Commission (EC) and the United States are to reopen talks after more than a decade aimed at eventually harmonizing US patent law with that of most other countries.

The talks are part of the Transatlantic Economic Partnership, the regular bilateral trade negotiations between the EC and the United States.

A first round is expected to be held in Washington DC in the coming weeks, to be followed by a second round in Brussels later this year. Informal talks could start as early as next week at a meeting of the World Trade Organization in Geneva.

Whereas priority on patents is attributed in Europe on a 'first to file' basis, the US system is based on 'first to invent'. The legal ambiguities this can create lead to expensive legal squabbles to establish the paternity of inventions. The United States is the only country to have such a system.

A further difference is that in Europe a patent has to be applied for before publication, whereas in the United States a 12-month post-publication grace period exists.

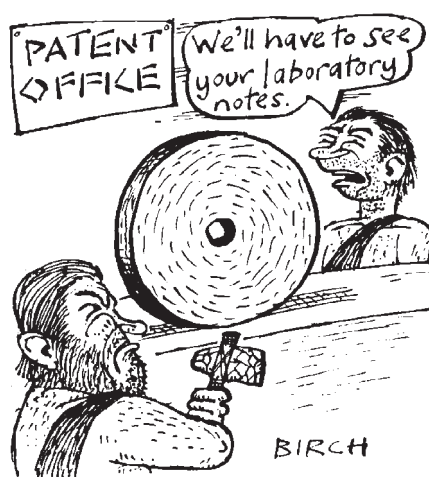
'First to file' has become the standard outside the United States because it is less complicated and costs less to process than applications based on first to invent, particularly if patent lawyers are employed by competing clients to prove that they were first to come up with a particular idea.

'First to file' is particularly popular with small companies and with lone inventors who, unaware of the requirements of US patent procedures, find that their attempts to patent discoveries in the United States run into difficulties if they lack well-kept laboratory notebooks that have been authenticated by witnesses.

The first-to-invent system is nonetheless staunchly defended in the United States by many small companies and inventors. William Respass, a senior vice-president of Ligand Pharmaceuticals in San Diego, California, argues that, under this system, the patent holder is genuinely the first person to come up with an idea.

The United States is unenthusiastic about harmonizing its patent laws with the rest of the world, partly because it believes them to be fairer to inventors and scientists by allowing them to publish their work before applying for a patent.

Harmonization was last attempted by the World Intellectual Property Organization a decade ago. But this slipped off the agenda when talks at the World Trade Organization led to the signing of the multilateral Trade



Related Aspects of Intellectual Property Rights (TRIPs) agreement.

An attempt to introduce first-to-file legislation in the United States in 1992 failed following strong opposition from small companies and inventors. Many large corporations, however, support a move to first-to-file law on the grounds that a single worldwide system will make the job of applying for patents easier.

An official at the EC in Brussels says that some progress is expected before the end of the year. But Dave Schmickel, patent and legal counsel for the Biotechnology Industry Organization in the United States, is one of many observers who believe that obtaining a change in US patent law could take "decades rather than years".

One leading US patent lawyer says that, if the United States were to move to a first-to-file system, inventors would expect to be

allowed a grace period, as well as a mechanism for protecting the theft of an idea.

The EC has considered whether to adopt a grace period. But the idea is not popular, and the commission is reluctant to go down this road, partly because establishing patent paternity could be a recipe for expensive litigation (see *Nature* 395, 531; 1998).

A compromise may also have to be found on the issue of transparency in processing a patent application. Unlike countries in the European Union, patents in the United States are processed behind closed doors, which allows experimental and other details to be hidden from competitors.

But Ron James, research director of PPL Therapeutics, the Edinburgh-based company that holds the patent for Dolly the cloned sheep, says that a transparent system allows others to monitor what might be fraudulent applications.

The talks are being backed by Britain and other European countries which are looking to small biotechnology and information-technology businesses to become major contributors to economic growth.

Nick Scott-Ram, chair of the intellectual property advisory committee of Britain's BioIndustry Association, welcomes the moves towards harmonization. He says an increasing number of small and medium-sized companies are now aware of US patent requirements, such as the need to keep accurate records during the process of discovery.

But Scott-Ram says that such companies are equally concerned that the talks should result in progress on reducing the costs of applying for a US patent. Ehsan Masood

German call to invest more in technology

[MUNICH] German companies risk an erosion of their 20 per cent share of the world pharmaceutical market unless they invest more in new technology, according to a new report.

The report — commissioned by the federal research ministry from six independent institutes for economic and innovation research — concludes that companies need to increase their investment in biochemistry, bioinformatics and combinatorial chemistry.

It also reveals that public and private funding of research in Germany is only now beginning to recover from large cuts made in the first half of the decade.

Support for research and development fell from around 3 per cent of gross

domestic product (GDP) in 1990 to 2.3 per cent in 1996. Although public funding has remained relatively flat, industry has increased R&D investment by 10 per cent between 1995 and 1997. Research funding is now 2.4 per cent of GDP, compared with 3.6 per cent in Sweden, 2.8 per cent in Korea and Finland, 2.7 per cent in Switzerland and 2.6 per cent in the United States.

The report, *Germany's Technological Capacity and Efficiency*, identifies the country's strengths as being in solid-state physics, semiconductors, material science, polymer research, measurement technology and astrophysics. But it still lags behind Britain, Japan and the United States in key technologies such as software development and biotechnology. Quirin Schiermeier

DoE budget scramble as closure plan axed

[WASHINGTON] The US Department of Energy's delicately balanced, \$320 million budget for nuclear physics was thrown into turmoil last week, just as it was being presented to Congress, when energy secretary Bill Richardson revoked its plan to close down the Bates Linear Accelerator. This facility, built in 1968, is operated by the Massachusetts Institute of Technology (MIT).

According to a spokesman for MIT, Richardson called the institute's president, Chuck Vest, on 1 February, the day the budget was released, to inform him that its provision to close Bates at the end of this year was a mistake. The following day, Martha Krebs, assistant secretary for science at the Department of Energy (DoE), issued a statement saying that Bates would be funded after all. The department is expected to issue a budget amendment within the next month, saying where it will find the extra \$10 million needed to keep the facility open.

Ernest Moniz, a former director of the Bates accelerator, is now under-secretary for energy, but is said to have excluded himself from the decision.

MIT plans to use some of the money to complete the Bates Large Acceptance Toroid (BLAST) experiment, which will study the basic physics of magnetism in nuclei. "I'm delighted that we'll be able to continue the work," says Robert Redwine, director of MIT's nuclear science laboratory.

But the decision leaves DoE staff scrambling to find the money without hurting the rest of the nuclear physics programme. Last



Reprieved: the \$10m needed to keep the Bates accelerator open will have to be found elsewhere.

year, a sub-panel of the department's Nuclear Science Advisory Committee (NSAC) recommended that, unless more money became available for nuclear physics, Bates should be closed to allow funds to be diverted to more modern medium-energy physics facilities, primarily the Continuous Electron Beam Accelerator Facility (CEBAF) at Newport News, Virginia. In last week's budget, nuclear physics received an overall increase of just \$4.5 million, to \$343 million.

"We're going to work very hard to have it

not impact the rest of the nuclear physics budget," says Krebs. "We'd like to see MIT make a contribution as well."

Claus-Konrad Gelbke of Michigan State University, the chairman of NSAC, says he welcomes the news that Bates will stay open. "We never felt there was a scientific justification to close it," he says, adding that closure had only been recommended if no additional money was available. "I hope there will be an increase in the nuclear physics budget — if not, they are robbing Peter to pay Paul."

Pressure from Senator Edward Kennedy (Massachusetts, Democrat) and Congressman John Tierney (Massachusetts, Democrat) may have led to Richardson's decision. The congressmen were concerned about job losses at the facility, which employs 85 people.

Officials involved in reallocating the funds warned that it would be difficult to do so without upsetting the fine balance that currently exists between different elements of the department's budget, and that the money for Bates can only come from other science programmes.

Herman Grunder, the director of CEBAF, was not available to comment, but a spokeswoman said that he had been happy with the budget of \$73.7 million which the facility was promised, before the Bates reprieve. CEBAF's budget may be more secure than the \$118 million that the DoE pledged last week for the first year of full operations at the Relativistic Heavy Ion Collider at the Brookhaven National Laboratory in New York state.

Colin Macilwain

MIT/BATES

UK panel formed to rebuild trust in government science advice

[LONDON] The British government has formed a group of experts to provide it with advice on communicating risk, in a bid to restore public confidence in the ability of government to handle issues such as food safety. Public faith has been shattered in particular by the crisis over bovine spongiform encephalopathy (BSE).

The 28-member group includes leading academics, journalists, heads of non-governmental organizations, and government officials. It includes Sir Robert May, the government's chief scientific adviser, Liam Donaldson, the chief medical officer, and Jon Snow, a prominent UK TV news presenter. Its deliberations will feed into a government seminar on risk management to be held in the coming weeks.

The group's creation coincides with the publication of a report from the UK Consumers' Association, which concludes that a science-based approach cannot alone be relied upon to reach socially acceptable decisions on issues involving the

communication of risk. The report, *Confronting Risk — A New Approach to Food Safety*, says: "Whilst science has an important role to play, it can only take us so far. It is important to recognize that scientific assessment itself is not a value or a judgement-free process. And even if it were, it is often surrounded by uncertainty."

The report concludes that the government needs to be more open and transparent in the way it manages risk, acknowledging scientific uncertainty, and involving as many relevant people — particularly from the public — as possible in the decision-making process.

It also recommends that meetings of scientific advisory committees should be held in public wherever possible, and that "minority scientific views" should be considered when picking experts to sit on these committees. The government is currently reviewing the structure and workings of its scientific advisory committees related to biotechnology.

A survey conducted last month by the polls company MORI revealed alarmingly low levels of public trust in government officials, including government scientists. Almost two-thirds of respondents "trusted most" independent university scientists and pressure groups — such as Greenpeace — to advise on the risks of pollution. Only 23 per cent trusted government scientists, and just 6 per cent ministers. Responses on the risks of BSE showed a similar pattern.

When asked to rate the government's handling of 13 issues, respondents placed genetically modified foods as the area handled least well. Modified foods topped the chart of issues where those polled felt that more legislation was required.

Sixteen per cent of respondents felt that they were well informed of the health risks of genetically modified foods, a similar proportion to those who felt informed about raw, unpasteurized milk. In contrast, 90 per cent felt well informed about the risks of smoking.

Ehsan Masood

Scientists fight for right to withhold data

[WASHINGTON] A law forcing scientists to hand over raw data under the Freedom of Information Act (FOIA) could be weakened by new proposals from the White House.

But the plan by the White House Office of Management and Budget (OMB) to interpret the law narrowly has done little to dampen resistance from universities and scientific societies, who want it revoked. They acknowledge the OMB move as an effort to take researchers' concerns into account. But they still fear the law would expose researchers to harassment from people who wish to interfere with their work, while making it difficult to protect confidential data related to medical studies or intellectual property claims.

The proposed rule, published on 4 February, explains how the OMB will implement a piece of legislation added to last October's Omnibus Appropriations bill by Senator Richard Shelby (Republican, Alabama). Shelby's law, incorporated at the last minute, requires non-government scientists on federal grants to hand over all raw data to third parties who request it under the FOIA.

The OMB rule, on the other hand, would only require raw data to be made available when the final research findings had already been published, and only if that research had been "used by the federal government in developing policy or rules". Comments are invited until 5 April, after which the OMB will issue a binding rule.

A 1980 Supreme Court ruling confirmed that university researchers and other people on grants are not subject to FOIA requests, which are mainly used as a means to obtain government documents. Scientists fear that the new law will expose researchers in controversial areas — such as global warming or the epidemiological study of the effects of pollutants — to demands from interest groups who want to see all their notebooks, papers and computer records.

Shelby had quietly inserted his legislation into the 4,000-page omnibus appropriations bill in response to a case in 1997, when industrial groups failed to persuade Harvard University to hand over the raw data behind a piece of published research cited by the Environmental Protection Agency in support of tightening regulations on fuel emissions. The study included medical details provided on condition of anonymity: researchers said that, even with names taken out, people could still be identified. A few other scientists have been allowed to use the data. But industry groups such as the American Petroleum Institute say that the data should be available for review by any qualified investigators with "a legitimate scientific interest".

"A researcher's notebooks are full of raw material, just like an FBI file, not all of which



Shelby: researchers united in opposition.

that is acceptable to the research community. But he adds: "I think the problem is that the basic statute is bad, and a lot of people believe that you can never make a bad law good." It is "very probable" that the association of medical colleges would now back legislation to repeal the Shelby legislation, he says. "We still think this is very troubling."

"We believe that OMB has made a good-faith effort, and has demonstrated sensitivity to our community's concerns," said Nils Hasselmo, president of the Association of American Universities, which represents the

is of equal standing," says David Korn, senior vice-president for research at the American Association of Medical Colleges. "To require all of that to be made publicly available is just bad science policy."

Korn describes the proposed White House rule as "a very valiant effort" to implement the law in a manner

leading US research universities. "We are consulting with our campuses and other scientific societies to provide extensive comments to OMB."

George Brown (California), the senior Democrat on the House Science Committee, has already proposed such legislation. But, says Korn: "Republican members are going to have to get on to this if it is going to go anywhere." Republican supporters of science such as John Porter (Illinois) have joined Brown in expressing their concern on the issue. But it will be awkward for them to propose the immediate repeal of a provision which was added to the appropriations bill by leading senators of their own party.

Shelby's office said that he had not yet reviewed the proposed rule.

Lawyers at the National Institutes of Health (NIH), meanwhile, have warned scientists that they will have to pay themselves for extra administrative work they do to satisfy demands made under the new law. NIH won't be able to pay for such work, they say, and even if NIH obtains a fee from the organization asking for the data, the money will go straight to the US Treasury. Colin Macilwain

SOIREE aims to iron out atmospheric puzzle

[LONDON] Oceanographers are trying to resolve a decade-long debate over the way iron levels in the oceans influence global climate. Researchers from six countries participating in the Southern Ocean Iron Release Experiment (SOIREE), led by New Zealand's National Institute of Water and Atmospheric Research, set sail from Wellington, New Zealand, last week on the research vessel *Tangaroa*.

The circulation and mixing of the Southern Ocean is such that the availability of iron to phytoplankton may have directly affected its ability to lock up atmospheric carbon dioxide. In the Southern Ocean, as in around one-third of the world's oceans, phytoplankton growth fails to make full use of available nitrate and phosphate nutrients. This suggests that other factors are limiting photosynthesis, and with it the ability of phytoplankton to fix carbon dioxide, an important greenhouse gas.

Iron is essential for photosynthesis, and a shortage has been shown to limit growth in the equatorial Pacific Ocean, where surface waters fertilized with iron 'bloomed' (see *Nature* 383, 495–501; 1996). This oceanic region seems to have been rich in iron in the geological past, but this probably had little effect on atmospheric carbon dioxide levels — the physical circulation would have prevented any extra carbon fixed being locked away in the deep ocean.



Into deep water: *Tangaroa* faces stormy weather as it tackles the climate-change issue.

In contrast, if phytoplankton productivity in the Southern Ocean is limited by a shortage of iron — the main question SOIREE aims to answer — an increase in iron levels may have significantly reduced the concentration of atmospheric carbon dioxide in the past. Much of the consequently fixed carbon in surface waters would be expected to be physically mixed down into deep waters, removing it from contact with the atmosphere for long enough to diminish atmospheric levels.

The major problem facing the expedition is the weather. Storms in the rough Southern Ocean will make it difficult to find the 10–15 day window of calm seas needed to maintain and follow the 50 square kilometres of surface ocean that will be seeded with ferrous sulphate. Philip Newton

Roche and Promega back in court for *Taq*

[SAN FRANCISCO] The seven-year battle over ownership of rights to the enzyme *Taq* DNA polymerase came back to court this month.

A San Francisco court will hear allegations that the US patent held by Swiss drug company Hoffmann-La Roche was fraudulently obtained by the original applicant, Cetus Corp. Roche is being challenged by Promega Corp, a laboratory supply company based in Madison, Wisconsin, which has been selling *Taq* since the late 1980s under a licence granted by Cetus.

The dispute dates from Roche's 1991

purchase of the rights to the polymerase chain reaction (PCR) from Cetus in a \$300 million deal. At the time, the Swiss company asked Promega to stop selling *Taq*, claiming it violated the Roche-Cetus agreement. When Promega refused, Roche demanded higher royalties. Promega again refused, and in 1992 Roche filed a suit for breach of contract. The contract expired in 1995.

In November 1996, Roche received a broad European patent covering key PCR enzymes including *Taq*. At the time, company officials said the approval showed that

expert scientific examiners recognized the biochemical facts that distinguish the Cetus invention from prior work. A year later, however, the Australian Patent Office withdrew rights to the enzyme in its native form, in response to a challenge led by New England Biolabs and Bresatec (now Bresagen) on the basis that other researchers knew of *Taq* before Cetus made its claim (see *Nature* 390, 327; 1997). The Australian Patent Office upheld Roche's claim to recombinant *Taq*.

Promega argues that Roche's *Taq* patent is invalid because *Taq* had been anticipated by earlier work. Its lawyers claim that Cetus was aware that the thermostable DNA polymerase it had extracted from *E. coli* was the same as enzymes from the same organism that had already been described in the literature, and so deliberately misled the US Patent and Trademark Office.

The US District Court Judge Vaughan Walker had ruled in 1996 that *Taq* inventors made four material misrepresentations or omissions in their patent application that could have misled patent examiners. Now he must decide whether those misrepresentations were intentional, and if Cetus scientists and executives knew their invention was not novel and withheld information.

Proceedings began on 1 February with Arthur Kornberg, winner of the 1959 Nobel Prize in Medicine, on the witness stand. As he explained the isolation of DNA polymerase, Judge Walker — in what has become a leitmotiv — urged participants to stick to the point. In cross-examination, Roche lawyers then showed that two of Kornberg's exhibits included erroneous details.

More cut and thrust can be expected in the two weeks scheduled for the trial. Using internal memos and expert testimony such as Kornberg's, Promega hopes to show that David Gelfand, the primary inventor on the *Taq* patent and manager of the *Taq* project at Cetus, misrepresented facts about *Taq* activity, nuclease contamination, measured fidelity, pH profile and other characteristics that Cetus used to show the invention was novel.

Promega plans to lay out 13 instances of misrepresentation or material omission in the patent filing. The documents and other information show that "the inventors chose to mislead, misrepresent and omit critical information time and again", writes attorney James Troupis in Promega's trial brief.

Roche responds that the applicants either did not know about the undisclosed information or did not believe it was important. The Cetus discoveries were distinct from earlier ones and the scientists involved are innocent of deception, writes Roche attorney Brian Poissant, adding that some exaggeration and advocacy in a patent claim does not amount to fraud.

Sally Lehrman

US sanctions hurt basic research in India

[NEW DELHI] Sanctions imposed by the US government after India's nuclear tests in May 1998 have badly hit work at the Tata Institute of Fundamental Research (TIFR) in Mumbai (formerly Bombay), India's premier institution for physical and mathematical research.

Some 45 of the institute's scientists held an extraordinary meeting on 28 January. They said the situation was "serious" and that some projects may be abandoned if the sanctions were not lifted soon. They believe other research institutes in India have similar problems.

TIFR, funded by the Department of Atomic Energy, is one of 250 "entities" the US government claimed were helping India's nuclear and missile programmes. US government agencies were prevented from collaborating with these, and US firms require export licences to trade with them (see *Nature* 393, 197 & 396, 206; 1998). Last month, the US government stopped seven scientists from the Fermi National Laboratory from attending an international conference on high-energy physics at TIFR.

Chemistry professor G. Krishnamoorthy, who organized the meeting, said that TIFR was suffering most from the denial of equipment and services from US firms. "The effects of sanctions began only after the publication of the entities list by the US Bureau of Export Administration in the middle of November," he said. "We have received a large number of denials from companies because they apparently want to play it safe [with the US government]."

A laser bought from Spectra-Physics at a cost of \$250,000 in March 1998 for studying biomolecules has been idle for months because the company is refusing to ship any attachments or to service it, despite a one-year warranty. A \$100,000 computer bought from Silicon Graphics "for solving complex problems in theoretical physics" is lying unused, as the company has refused to replace defective processors. Work stations



Hit parade: the Agni missile makes its New Delhi debut in last month's Republic Day celebrations.

and servers bought from Digital are not being serviced. Sorvall has refused to deliver high-speed ultracentrifuges. Malaria researcher Shobhona Sharma says that Sigma (India) has been instructed not to supply chemicals or biochemicals. Optical and electronic components from Thorn Labs have been denied and, says Krishnamoorthy, "this is only a sample".

Scientists say that research using equipment supplied before the sanctions and not working properly has stopped. "Where new purchases of materials and equipment have been denied, the ongoing research projects have either had to be abandoned or toned down," says Krishnamoorthy. Although some items can be obtained from other countries, he says, this is impossible in many cases. "Denial of products from the USA would place us in a situation where we will not be able to select the best approach in our research."

TIFR scientists feel the sanctions have "gone far beyond the stated purpose of limiting India's nuclear, missile and military activity" and are hurting basic science. Some suspect that the heads of the country's scientific agencies are downplaying the issue because of vested interests.

TIFR scientists fear that if this continues "we may not be able to attract bright students and scientists" — a serious problem since many of its scientists are due to retire in the next few years. K. S. Jayaraman

German science fights animal rights bill

[MUNICH] Proposals to modify Germany's constitution to include a provision protecting the rights of animals have come under fire from the country's main research agencies, as well as pharmaceutical and chemical companies.

The change is being proposed by the governing coalition of Social Democrat and Green parties. The addition would state that animals have the right to be "respected as fellow creatures" who have an individual right to be protected from "avoidable pain".

Similar changes have also been proposed by two opposition parties: the Free Democrat Party and the reformed East German Communist Party.

Vocal resistance to the change has come from the research community, with opposition expressed by Germany's largest research council, the Deutsche Forschungsgemeinschaft (DFG), the Max Planck Society, the Conference of University Rectors, and the Conference of Mathematical and Scientific Faculties. They argue that it would damage biomedical research — a position supported by the Association of the Chemical Industry, and the pharmaceutical company BASF.

In a statement, the DFG questioned the need to change the constitution, saying that Germany's animal rights legislation is among the strictest in Europe. They also claimed that such a provision would conflict with the constitutional protection of research freedom.

"If animal rights become a constitutional issue then the scope for legal challenges to experiments could be immense," says Hubert Markl, president of the Max Planck Society. This could result in approval procedures being delayed for "months or even years", he claims. "Scientists in highly competitive fields would either change the direction of their research or move elsewhere."

The Conference of University Rectors warned that constitutional protection of animal rights risks legitimizing and encouraging the activities of militant groups opposed to animal research (see *Nature* 396, 505; 1998). And one observer argues that in practice such a measure would do little to alleviate animal suffering through intensive farming and long-distance livestock transport.

At a first reading in the Bundestag, Germany's federal parliament, last month, the moves were challenged by speakers from the main opposition party, the Christian Democrats (CDU), who backed scientists' arguments that such a change could damage the country's biomedical research base.

Without support from the Christian Democrats, this constitutional change is unlikely to obtain the two-thirds majority vote required at the final second hearing.

A question mark also hangs over the support for the proposal within the Social Democrat party itself. The chancellor, Gerhard Schröder, has expressed commitment



Laboratory life: Germany proposes to give animals constitutional rights as individuals.

to such a change, but observers point out that this was before the party came to power, and that a unanimous 'yes' vote from Social Democrats is unlikely.

In a bid to attract support within the party and from the Christian Democrats, the government is considering a watered-down proposal which would affirm the need to protect animals, but would stop short of giving rights to individual animals. *Quirin Schiermeier*

Japanese university approves genetic tests on *in vitro* embryos

[TOKYO] The first clinical application of preimplantation genetic diagnosis in Japan moved a step closer last week with the approval by the ethics panel of Kagoshima University Medical School of a protocol for testing for Duchenne muscular dystrophy in human eggs fertilized *in vitro*.

A decision on whether to go ahead depends on the approval being confirmed by the Japan Society of Obstetrics and Gynaecology. The society last year opened the way to preimplantation genetic diagnosis in principle when it approved general guidelines covering such procedures (see *Nature* 394, 110; 1998). These allow couples to ask for eggs used for *in vitro* fertilization to be discarded if susceptibility to certain types of hereditary disorder is detected in the genes.

But the society came under fire, as its guidelines were seen as a poor replacement for a national debate on the ethical issues. Critics argued that legislation was needed to prevent such techniques being put to non-medical uses, such as selecting the sex of a child.

Akira Hira, principal of the university's

medical school and chairman of the ethics committee, says its decision was based on the safety and reliability of the technique, and the potential benefit it would bring to patients. But he admits that gaining public acceptance will not be easy. "There are still a lot of ethical and scientific questions we would have to explain to the public to gain their support and understanding," he says.

Concern about the ethical implications of Kagoshima University's move has been voiced by the anti-eugenics patients' advocacy group, which represents more than 30 women's and disabled groups. They argue that prenatal diagnosis risks increasing discrimination against the disabled.

But Hiroyuki Nagata, professor of reproductive medicine at Kagoshima University, argues that early detection of genetic disorders would allow mothers undergoing *in vitro* fertilization to avoid 'therapeutic abortions'.

The approved protocol involves a couple who already have a child suffering from Duchenne muscular dystrophy. "I feel that

the protocol poses fewer ethical problems than prenatal genetic diagnosis [of fetuses] and therapeutic abortion does," says Nagata.

The university originally approved the same protocol in 1995, but was forced to postpone a final decision because of opposition from patients' advocacy groups. Since then, the society has issued guidelines, and the latest proposals are based on them.

Opposition from patients' advocacy groups has similarly forced the Japan Society of Obstetrics and Gynaecology to postpone approval of specific guidelines for screening for Duchenne muscular dystrophy (see *Nature* 385, 763; 1997). The general guidelines approved last year do not specify which diseases can be tested for, but limit screening to serious diseases, and require case by case approval by a specialist committee of doctors and geneticists appointed by the society.

Keiko Yano, the director of the advocacy group, argues, however, that the ethical issues the technique raises require greater discussion before any further decision is made on the application of preimplantation genetic diagnosis. *Asako Saegusa*

Everglades plan flawed, claim ecologists

[WASHINGTON] Leading ecologists have written to US interior secretary Bruce Babbitt, warning that an ambitious and politically prominent \$8 billion scheme to restore the Florida Everglades has "serious failings" and needs independent scientific review.

"These are deep, systemic problems, unlikely to be overcome by tinkering with the existing alternative," said the 28 January letter, signed by Stuart Pimm of the University of Tennessee, Paul Ehrlich of Stanford, E.O. Wilson of Harvard, Gary Meffe of the University of Florida, Peter Raven of the Missouri Botanical Garden and Gordon Orians of the University of Washington. Babbitt has agreed to hear the scientists' concerns.

The scheme calls for the US government and the state of Florida to share the costs of refurbishing the region's dykes and canals, digging storage reservoirs, restoring the ecosystem and providing fresh water and flood protection for the booming south Florida population. The restoration is expected to take at least 20 years.

The project aims to reverse the legacy of half a century of diversion and disturbance of the region's natural flow of water. Soil erosion is rife, the water table is dropping, flooding is frequent, rural and urban runoff have polluted local waters and habitat loss threatens biodiversity. The original Everglades have shrunk by half, while the south Florida population has grown from two million in 1948 to six million, and is estimated to reach 15 million by 2050.

But Pimm, who drafted the letter, and has worked extensively in the Everglades, questions the scheme's reliance on a computer model of Everglades hydrology. He claims the simulation is unrealistic, and that, for example, adjacent marsh areas with vastly different flooding periods do not occur in nature.



In the mire: the goals and methods of the Everglades restoration plan are under attack.

The letter also claims that repeated calls from National Park Service (NPS) and US Geological Survey scientists to explore natural solutions to managing the region's water rather than technological fixes have not been given sufficient consideration.

Government planners and environmentalists have responded angrily to the claims. "An apology is neither required or appropriate at this time," said Michael Davis, a deputy assistant secretary in the US Army, in a letter to Pimm last week. Davis, whose Corps of Engineers is co-directing the Everglades 'restudy', denied it was flawed: "I have the utmost confidence in the interdisciplinary, interagency team that developed the plan."

The letter is dismissed as "minority opinion" by David Guggenheim of the Conservancy of Southwest Florida, who chairs an

Everglades Coalition of more than 30 environmental groups. Mainstream environmentalists "all feel that this has been a good scientific process," he says.

This late questioning of the project's scientific integrity is "not appropriate," says Guggenheim. Scientists and engineers have worked for nearly 10 years on the restudy, and the plan is scheduled to reach Congress in July.

Pimm, speaking personally, alleges that the plan places water supplies above restoring the ecosystem. He cites NPS scientists who, in a critique of the restudy just before Christmas, said "There is insufficient evidence to substantiate the claims that [the plan] will result in recovery of a healthy, sustainable ecosystem. Rather, we find substantial, credible evidence to the contrary."

"This is pretty damning," says Pimm, arguing that the NPS has been ignored. Robert Johnson, a scientist with the Everglades National Park research centre that produced the report, says only that there are "some outstanding issues" with the restudy.

Environmental recovery is a key goal, say advocates of the restudy, but this will happen more slowly than other goals of the scheme. "It will take many years to restore this important ecosystem," wrote Davis.

The scientists' letter to Babbitt stresses the need for an independent outside review of the scheme. "The National Research Council [NRC] would be an obvious choice," they wrote. The scheme includes plans for scientific oversight, but not until the work is under way. "Peer review is an integral part of our implementation strategy," says Davis.

Guggenheim says additional peer review is unnecessary. He calls the restudy an "exceptionally open" project, and says scientists from government agencies and green groups have been looking over the restudy managers' shoulders throughout the planning phase. "We're not pushovers," he says of the Everglades Coalition; "we're very critical". The Sierra Club has broken ranks with other environmental groups, however, by echoing the call for outside peer review.

The South Florida Ecosystem Restoration Task Force, which includes representatives of various participating agencies and organizations, issued a statement last week saying it "welcomes scientific review of the restudy from all quarters at all times."

A formal NRC study could take years. The prospect of a delay worries project supporters, who have only recently got the political forces in southern Florida "all singing from the same page," says Guggenheim. He says a hold up may send "all the wrong signals. If we do anything to delay delivery of this [restudy] to Congress in July, it just might not happen."

Tony Reichhardt

Chandra slips, space shuttle runs a risk

[WASHINGTON] The launch date for the US\$2 billion Chandra X-Ray Observatory, originally scheduled for August 1998, was set back from May to July 1999 last week, the fourth postponement in just over a year.

Most previous delays had involved problems with the spacecraft, but the new delay has been caused by space shuttle missions being rescheduled to launch part of the International Space Station in May.

The delays to Chandra, third of the "Great Observatories", have cost NASA some \$50 million, but have not otherwise compromised the mission, say project scientists. The observatory, renamed to honour the late Indian-American astrophysicist Subrahmanyan

Chandrasekhar, will operate until 2004.

NASA had more bad news last week, from its Aerospace Safety Advisory Panel. The panel said in its annual report that personnel cuts in the agency and the shuttle's private operator, the United Space Alliance, have jeopardized safety to the point where NASA is "moving toward losing the core competencies needed to conduct the nation's space flight and aerospace programs in a safe and effective manner".

It advised the agency to ensure adequate budgets for its three field centres involved in human spaceflight. The report suggests that NASA's tactic of subsidizing science spending by cutting the shuttle programme may finally have come to breaking point. T.R

French ministry in climbdown over reform

[PARIS] The French ministry of national education, research and technology has quietly abandoned plans for a decree to reform the statutes of the country's main research agency, the Centre National de la Recherche Scientifique (CNRS).

The ministry, while officially continuing to refuse researchers' demands for a national debate, has in parallel now requested a parliamentary inquiry into the mobility of researchers and into ways of improving links between the public research agencies and the universities.

The decree to reform the CNRS has been a major focus for strong opposition in the research community to plans by Claude Allègre, the science minister, for a profound reform of the country's research system (see *Nature* 396, 607 1998). Scientists complain that it would transfer excessive responsibility for the work of CNRS laboratories to the universities, while failing to address fundamental problems, such as the weakness of university research.

Vincent Courtillot, a former principal adviser to Allègre who was recently appointed director general of the ministry, discreet-

ly announced the abandonment of the CNRS decree last week in an interview with the newspaper *Le Monde*. The format of a decree was too rigid, he said, adding that reform of the CNRS would await proposals from the organization itself "within three to four months".

The conciliatory tone of Courtillot's policy outline, in sharp contrast to the previous hardline stance of the ministry, has been interpreted by observers as indicating that the ministry is keen to seek a way out of the current deadlock over reforms. "It marks a substantial change on the ministry's part," says one member of the CNRS board of directors.

The deadlock culminated in an unprecedented meeting in Paris just before Christmas of the 800-member National Committee for Scientific Research, the 'parliament' of the country's scientists, which plays a major role in evaluating laboratories and administering recruitment. The meeting attacked the reforms as "ill conceived" and overwhelmingly rejected the way the ministry has tried to impose them on the scientific community with minimum consultation.

This month, the presidents of the 40 sections of the national committee issued a statement arguing that a national debate was essential. Courtillot maintains that such a debate is not on the cards, arguing that, had the government wanted one, it would have organized it on coming into power 18 months ago.

But one member of the CNRS board of directors says this amounts to "saving face". Indeed, the ministry has asked for a parliamentary inquiry into research issues that is expected to lead to concrete proposals. The ministry has recently come under fire from its own ranks, with the research committee of the ruling Socialist party supporting calls for wider consultation between the ministry and the scientific community.

Depending on the inquiry's remit, it might well satisfy researchers' demands, says Henri Edouard Audier, a member of the national board of SNES, the main trade union representing researchers.

At the same time, Audier laments that a procedure of consultation could have already been completed had it been under way from the outset.

Declan Butler

British universities face review to increase accountability

[LONDON] A review of research spending in British universities to make them more accountable is to start this year in a series of pilot institutions. The so-called 'transparency review' will be extended to the top research universities by the end of next year.

The move was announced last week by John Taylor, the new director-general of the research councils. Taylor was speaking before the House of Commons Science and Technology Select Committee in his first major policy statement since taking office six weeks ago.

He also indicated that he questioned the need at this time for a new government white paper (policy document) on science — an idea floated at a recent meeting of politicians and scientists in Downing Street (see *Nature* 396, 714; 1998). "There are no immediate issues staring us in the face," he said.

The transparency review was a condition attached to increases in research spending allocated to universities following the government's recent comprehensive spending review (see *Nature* 394, 209; 1998). The review will be overseen by the Science and Engineering Base Co-ordinating Committee, chaired by the government's chief scientific adviser, Sir Robert May.

The main aim of the exercise is to make



Taylor: focusing on identifying costs.

universities more accountable for research spending and to oblige them to better identify research costs — this is 'activity based costing'.

At the moment, the universities receive core funding from the higher-education funding councils and project funding from the research councils, under what is known as the dual-

support system.

But in a surprise move, Taylor also said that the review will include an assessment of whether current funding arrangements are "sufficiently selective", with the distribution of funding resulting in the "identification and maintenance of centres of excellence". The review will also consider whether these two streams of government research funding are "complementary and coherent".

The transparency review will initially focus on the complex administrative issue of introducing an activity-costing methodology to universities. Taylor told the committee that final specifications for this should be available by June. It would then be

implemented in volunteer universities as early as the start of the next academic year and extended to all the top research universities towards the end of 2000.

Taylor explained to members of parliament that he had "pushed hard" to move the transparency review this far forward in the six weeks he had been director-general. He warned the committee that it would not be possible to bring it any further forward, saying: "You have to keep the aeroplane running while you change the engine." The review, he said, would bring "information that will help us better run dual support, and triple support if we have a third leg of funding. We are going to get a better understanding of whether there really is a funding gap." Under the present arrangements, universities do not receive the full costs for the research they undertake.

Taylor spelt out as a priority for his term in office the interface between the research councils and how well they coordinate their activities on key topics and interdisciplinary issues. He warned the committee that the excitement over genome research should not be allowed to lead to neglect of other areas of research, such as information technology and communications, which he predicted would dwarf other fields in terms of economic returns over the coming decade.

Natasha Loder

Bordogna backed by Clinton for NSF deputy director post

[WASHINGTON] Joe Bordogna, an electrical engineer who has been acting as deputy director of the US National Science Foundation (NSF) since 1996, has finally been nominated by President Bill Clinton to fill the post on a permanent basis.

The nomination will have to be confirmed by the US Senate. The deputy director of NSF is the chief operating officer of the agency, and traditionally takes much responsibility for its day-to-day operations. Bordogna has already been leading the agency's efforts to comply with the Government Performance and Review Act.

In a separate development, Nick Smith (Republican, Michigan), a Congressman who has had little previous involvement with science issues, has joined the Science Committee of the House of Representatives and been named as chair of its basic research subcommittee, which oversees the NSF.

Spain joins search for extraterrestrial life

[MADRID] **Daniel Goldin, director of the US space agency NASA, has announced that**

Spain is to host a new astrobiology centre within the National Technological Aerospace Institute, which is part of the Ministry of Defence.

There will be no funding from the United States, but the centre will be a NASA-affiliated institute. The Spanish government will provide initial funding estimated at 1,050 million pesetas (\$7 million).

"We don't know if life exists beyond our planet within the Universe, but we would like to find out," said Goldin. "We welcome Spanish scientists who will be working with us." The project will involve an interdisciplinary team of geologists, biologists, physicists and engineers.

Neutron source aids Russia's physicists

[MOSCOW] Russia's first high-intensity impulse source of neutrons was opened last week at the nuclear research institute of the Russian Academy of Sciences in Troitsk, south of Moscow.

The total cost will be 180 million rubles (US\$8 million), of which the academy and the ministry of science and technology will each pay 60 million rubles, with the rest coming from the institute itself. Vladimir Bulgak, deputy prime minister responsible for science, said the source will allow Russian

scientists to conduct fundamental and applied research to meet modern demands.

Bulgak also said the cabinet had allocated science 2.02 per cent of all budget expenditure in 1999, compared with 1.49 per cent in 1998.

Japan to relax guidelines on gene therapy

[TOKYO] Japan's Ministry of Health and Welfare is to revise its guidelines on gene therapy to allow its use to treat chronic diseases of blood vessels, nerves and muscles.

According to the ministry's announcement last week, the guidelines, which are expected to be revised by the autumn, will allow diseases such as arteriosclerosis and rheumatoid arthritis caused by blood vessel disorders to be treated by gene therapy. Under the current guidelines, gene therapy can be used only to treat serious life-threatening diseases such as cancer and AIDS. The revision will be accompanied by changes in guidelines on gene therapy research at universities.

Partnership scheme to boost Russian teaching

[MOSCOW] The Russian government and the Open Society Institute have launched a joint

project to improve the standard of education in rural universities and schools. The project has a budget of US\$15 million over the next three years.

Leading Moscow and St Petersburg educational and scientific organizations responsible for postgraduate teaching will be chosen as 'resource centres'. They will identify current or potential areas of growth in provincial universities, and provide them with sufficient support to compete with the best universities in the country.

Andrei Kortunov, the project's executive director, says: "We hope that foreign donors will join our project, especially if western universities are chosen as the resource centres."

US physics institute wins court victory over survey

[LONDON] The American Institute of Physics has won a case brought against it by Gordon & Breach Science Publishers and affiliates. The US Court of Appeals for the Second Circuit unanimously affirmed a previous decision by a district judge that the publication and promotional use of a survey of journal prices did not constitute false or misleading advertising.

The survey of journal prices was conducted by Henry Barschall and published

by the institute, and showed that Gordon & Breach's physics journals were, on average, far more expensive in terms of cost per thousand characters than those of other publishers.

Marc Brodsky, the institute's executive director, said: "We are extraordinarily pleased with the outcome because it allows the free flow of information that bears on the difficult problems that libraries confront in dealing with the rising costs of journals."

Genome project aims to combat prawn scourge

[TOKYO] A joint venture company established by three US companies in Shanghai has reached agreement with Chinese government agencies to fully sequence the genome of a baculovirus that has devastated prawn farms in much of the Asia-Pacific region.

Shanghai GeneCore BioTechnologies, a joint venture of PE Biosystems, Axys Pharmaceutical and SiniWest Holdings of the United States, will join hands with the China National Center of Biotechnology Development, the Beijing North Genome Research Center, and the Third Institute of Oceanography in Xiamen.

The aim is to sequence genomic DNA and cDNA of C-type baculovirus, commonly

known as prawn white spot baculovirus.

The virus has reached epidemic proportions. In 1993 alone it caused an estimated loss of US\$1.2 billion to China's prawn fisheries, which are the largest in the world.

Tory U-turn on genetically modified crops

[LONDON] Britain's main opposition party, the Conservative Party, has added its voice to the crescendo of calls seeking a moratorium on the introduction of genetically modified crops in Britain.

William Hague, leader of the opposition, called on the prime minister, Tony Blair, to take the advice of English Nature, the government's advisory body on wildlife and countryside issues, and delay plans for the commercial planting of modified crops until research on ecological risks has been completed. The move represents a significant shift in Conservative policy. The party is usually considered as more business friendly than ecologically minded.

The government and industry have agreed on a one-year moratorium on herbicide-resistant crops and a three-year moratorium on insect-tolerant crops. Field trials assessing the ecological risks of each are expected to last at least three years.

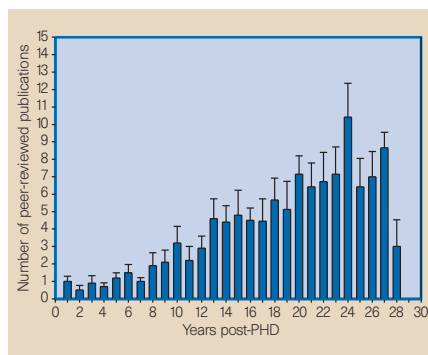
Pressure to publish stifles young talent

Sir — Almost three years have passed since the last government Research Assessment Exercise (RAE) conducted on UK universities. The next exercise will take place in 2001 and many universities are planning for the future with some anxiety. The consequences for departments that fail to retain ratings of 4 or 5 on the 5-point scale are dire. A lower score would mean loss of funding that might lead to redundancies. Meanwhile departments with ratings of 2 or 3 are already fighting for survival. All universities are trying to reach set RAE targets. Some plans are realistic, others may not be. Either way the pressure is rising.

One obvious strategy is to remove staff less active in research, encourage early retirements and replace with new, probably younger, academic staff. However, recruitment has to be carried out with increasing haste. The job requirements for this new generation of academic staff seem overwhelming and many post-doctoral researchers may feel inadequately prepared.

The RAE committees in the laboratory science areas require four excellent peer-reviewed publications. These will be the basic requirement for posts in the universities. Is that really beyond the capability of most young researchers?

To test this hypothesis, I retrieved from the BIDS database the publication record of ten scientists in my own field of molecular microbiology, who have been based throughout their entire careers in the UK.



Building to a peak: would these high achievers have made it through the present system?

They are now in very senior positions in universities or research institutes; their careers span a total of 262 years. They have all achieved worldwide status and respect. The question I posed was: "How would they have matched the criteria for RAE in the early stages of their careers?"

The graph above shows mean publication rate (with standard error indicated by bars) per year after completing their PhDs. The data were collected from the BIDS database as publications back to 1981. I had to rely upon citations for my information before this time. Despite this handicap, it was clear that all the scientists produced a number of cited publications, often citing their own research in publications after 1981.

The pattern of publication output over their careers is revealing. Generally they

produced few papers over the first ten years, although the quality of the work was clearly outstanding. During that time, each one went for an average of 3.8 years without apparently publishing (or producing a cited paper) at all. In one case there was a five-year gap! The conclusion I reached from this simple exercise is that application of the current RAE exercise might have precluded all of them from appointments to academic posts.

These scientists went on to do great things. They have initiated novel, sometimes risky research and built up new avenues of endeavour along with highly respected research groups. Indeed, their later publications reflect the increasing number of collaborations that they have generated.

The careful nurturing of young talent has had its rewards for UK science. The atmosphere that surrounds the RAE may be stifling the talented, either by reducing the urge to take imaginative risks in new areas or by simply driving them out of science. Those who take up the challenge may find the pressure to perform very hard to bear. The management of research in universities may become tyrannies that allow little time for thinking, teaching, home life or young families. One thing is certain: without committed and imaginative young scientists who enjoy the challenge of discovery, the future of UK science will not be fruitful.

Michael J. Larkin

The Quesor Centre, The Queen's University Belfast, Stranmillis Road, Belfast BT9 5AG, UK

Language bias discredits the peer-review system

Sir — The system of refereeing original articles in science is based on what is called "peer judgement". The term "peer" means "equal in rank and abilities", but this is far from reality. Journal referees are usually experts in their fields, meaning that most of us are refereed by our superiors. Their criticism is usually very constructive.

There is, however, a point on which the peer-review system is discriminatory: language. Almost all the referees' comments mention style. They usually say that the manuscript must be revised by a native English speaker. For many researchers, English is a second language learned in adulthood. This prevents us having the same command as an English speaker.

I write my articles in English, then I pay a native English speaker (British or American, usually with a university science degree) to correct the spelling, grammar and style. On

one occasion, I asked an English friend to correct one of my papers. He is a professor at Oxford University, a well known expert in his field with more than 250 papers published, and is editor-in-chief of a reputed journal. He did an intense revision of the manuscript. In spite of this, one of the referees said the manuscript should be revised by a native English speaker, without any specific criticism about the style.

What can I say? I guess for this referee, any style different from his own is inadequate. That kind of comment makes editors send a paper back to the author for revision. That is a waste of money, time and morale, especially when there is nothing to correct. I recommend referees to be more aware of this fact when judging manuscripts from non-English-speaking countries.

I have thought of a possible solution: the creation of what I have tentatively called an Institute for Correct English Style (ICES). Through this, qualified people would ensure that the paper is written correctly and the institute would issue a certificate of competence in English. A paper corrected by an

ICES member would not need referees' judgements on style. Reviewers would then focus on their real expertise: science. The peer-review system would again be equal.

Under the present system, if *Romeo and Juliet* had been written by González instead of Shakespeare, this great work would have been rejected.

Antonio J. Herrera

Departamento de Bioquímica, Bromatología y Toxicología, Facultad de Farmacia, Universidad de Sevilla, Spain

German job security leads to stagnation ...

Sir — Your editorial and News report touch on fundamental problems that have beset German universities for a long time (*Nature* **396**, 393 & 396; 1998). The system is rigid and bureaucratic with little incentive for efforts beyond minimal requirements.

Many professors strive to achieve high standards against a gradient of adverse

conditions dictated by federal and state university laws, compounded by legal systems that rule on salaries for technical and academic staff. The result is a stable system, but one that favours the mediocre. Salaries depend solely on age, family status, and years in the job.

After five years at the university, everyone attains the right to permanent, tenured employment almost regardless of quality. As such positions are usually already filled and are no longer created, a well trained technician or academic staff member has to leave even if funds for their salary are available from research grants and they do not necessarily seek permanent employment.

Compared to the United States, tenure is reached early and in high numbers. This creates complacency. Furthermore, no university in Germany (except a few private institutions) can select their students. They have no influence on selection criteria or numbers admitted, and no student pays a single Pfennig for tuition. No wonder that calls for payment by results are opposed.

But at least some faculties have begun to reallocate their scarce resources on the basis of objective criteria that allow the assessment of research results.

Eberhard Passarge

Institut für Humangenetik, Universitätsklinikum Essen, D-45122 Essen, Germany

... and reduces students' chances of publishing

Sir — It is essential to introduce pay by results and to abolish civil servant status for university teachers in Germany. Civil servant status — commonly known as a synonym of inefficiency — prevents the firing of academics who can't compete in the 'publish or perish' world of science. Academics should face competition, in the same way as every other part of the labour market does.

During my studies in Germany I observed that the lack of an efficient system for the evaluation of university teachers leads to a growing class of academics who know that their performance does not affect their income. (I do not want to attack those who accept the rules of competition and know that their reputation depends on the quality of their publications.)

Lecturers usually neglect their teaching duties, and deliver the same lectures unchanged for years. From my experience I know that students with less renowned supervisors have to work harder to have research published, and often do not get a job in industry because in a tough job market the reputation of the supervisor is essential. Their work is not assessed fairly

and they won't get the chance to prove their abilities. Because such incompetent supervisors waste time on inefficient administration, they cannot provide the necessary support to students. They lack the skills to submit applications for funds, so their students have to finish their work on state benefits. But such supervisors can't lose their posts.

I do not support the call to abandon *Habilitation*, the postdoctoral qualification required to become a member of the teaching staff. *Habilitation* is necessary for a fair judgement of the individual's skills as a leader and manager.

For the sake of good research and education in Germany, the introduction of competition at universities is overdue. Payment by results only threatens the incompetent. Students are evaluated every day — so why not their teachers?

Andreas Herde

Dyson Perrins Laboratory, Oxford OX1 3QY, UK

Frémiet's phantasms

Sir — As a one-time 'fledgling palaeontologist' inspired by Frémiet's sculptures in the Paris Jardin des Plantes, I read with interest Martin Kemp's piece on their relevance to the public understanding of palaeontology and evolution (*Nature*, **396**, 727; 1998). However, his interpretation prompts some comments.

Besides the bronze relief *Man Triumphant over Two Bears*, shown by Kemp, there is another sculpture by Frémiet in the Jardin des Plantes depicting the struggle between man and bear — with apparently a completely different outcome. *Le dénicheur d'oursons* ('Bear Cub



Darwinism or symbolism? Emmanuel Frémiet's Bear Cub Snatcher, Jardin des Plantes, Paris.

Snatcher) is a large bronze statue standing near a children's playground not far from the Galerie d'Anatomie Comparée et de Paléontologie. It shows a prehistoric hunter, a strangled bear cub tied to his waist, in the deadly embrace of an adult bear (presumably the mother). There is a knife stuck in the bear's throat, but the animal is clearly in the process of breaking the hunter's back. This statue is thus a 'mirror image' of man triumphant. So, even in the popular scientific mythology of the late nineteenth century illustrated by Frémiet's work, the triumph of man over beast was far from assured (as also in Frémiet's *Orang-utan strangling a native of Borneo*, also shown by Kemp).

But the prominence of Frémiet's depiction of the struggle between primitive man and beasts in the Galerie d'Anatomie Comparée et de Paléontologie is actually something of a paradox. Kemp's labelling of Albert Gaudry, who conceived the Galerie, as "the leading French advocate of Darwin's theories" is misleading.

Gaudry was an evolutionist, and wrote that he had read Darwin's *Origin of Species* "with passionate admiration" and had "savoured it slowly, as one drinks a delicious liqueur". Nonetheless he had no taste for natural selection as envisioned by Darwin, and admitted that he was "far from Darwin's philosophical ideas in some respects". (Gaudry's quotations here are translated from ref. 1).

Because of his strong religious belief in a harmonious Creation, in which chance and struggle had no place, Gaudry developed an idyllic view of evolution², in which carnivores fed on herbivores to put an end to their sufferings when they grew old, and living beings developed according to a benevolent divine plan. He was convinced that "there was no competition for life, everything was harmonious". This was a far cry from mainstream Darwinism, especially its nineteenth-century incarnation.

Seen in that light, Frémiet's work stands as a symbol for the triumph of the human spirit (seen by Gaudry as the "marvel of Creation") over brute strength rather than a depiction of a Darwinian struggle for life.

In any case, whatever their exact cultural significance, these late nineteenth-century sculptures are more artistically interesting than the derelict and scientifically inaccurate fibreglass stegosaur, a late twentieth-century addition, that now 'adorns' the grounds of the Galerie d'Anatomie Comparée et de Paléontologie.

Eric Buffetaut

*Centre National de la Recherche Scientifique
16 cour du Liéqat, 75013 Paris, France*

1. Gaudry, A. *Les ancêtres de nos animaux dans les temps géologiques* (Baillière, Paris, 1888).

2. Buffetaut, E. *A Short History of Vertebrate Palaeontology* (Croom Helm, London, 1987).

Physics takes the biscuit

Why did a light-hearted experiment attract so much attention from the media? The episode is an interesting lesson for those wanting to explain science to the wider public — equations do not always scare people away.

Len Fisher

Scientists wanting to share their picture of the world with a wider audience have a familiar problem — the knowledge gap. One doesn't need to be a writer to read and understand a novel, or know how to paint before being able to appreciate a picture. Some knowledge of what science is about, though, is a prerequisite for both understanding and appreciation.

Our intended audiences can often be trained on the spot, if we can persuade them to stay around for long enough. One effective approach is to use the 'science of the familiar', exemplified by my recent exercise on the physics of biscuit (or cookie) dunking. The project was a publicity agent's idea, and the results were presented at a press conference in London on "National Biscuit Dunking Day". Local publicity, fuelled initially by the eccentricity of the story, rapidly spread worldwide — I am still receiving requests for radio and TV interviews from countries as far away as Australia and South Africa.

What gave the story such global appeal? Even the Nobel prizes don't receive such coverage. Yet this minor bit of science is on everyone's lips, metaphorically speaking. Its success, I believe, reflects a hunger for accessible science. The lessons from this success, some of them surprising, are important for those who wish to share more serious science with a wide audience.

What is dunking? It is simply dipping your biscuit, cookie, doughnut or pastry into a drink such as tea or coffee, a process shown by chemists at Firmenich (Switzerland) to enhance flavour release by up to ten times. Although scientifically acceptable, dunking is often socially frowned upon, which is probably part of its attraction.

The physics of dunking is straightforward. A biscuit is porous, with interconnecting hollow channels between the crumbs. When the biscuit is dunked, capillary action draws the liquid into these channels: a similar process occurs when a piece of blotting paper is dipped into ink, or when ring stains form from dried liquid drops¹.

The problem for serious dunkers is that the wetted part of the biscuit becomes very soft, especially when the tea or coffee is hot. A biscuit is basically dried-up starch grains glued together with sugar: the hot liquid swells and softens the starch grains and dissolves the sugar. The wetted biscuit eventually becomes so soft that it collapses under its own weight.

The physicist's answer to this problem is

to dunk the biscuit so that part of it can remain dry (and mechanically strong) and support the weight of the wet bit. Hence, instead of holding the biscuit vertically when dunking, a physicist grips it at the edge, sliding it into the tea or coffee at a shallow angle, so that the lower surface is wetted but the upper surface remains dry. This explanation is so simple that I was able to talk radio interviewers through it, and have them perform experiments on air. Yet simplicity alone cannot explain the high degree of interest.

I decided to ask the interviewers themselves why they found the story so interesting. An important criterion for most was that the topic is slightly daring, but the scientific gloss added an objectivity that legitimized public discussion. Most interviewers confessed to a strong interest in science, coupled with a fear of looking foolish when asking questions. This fear barrier is much lower when discussing the science of the familiar, as the questioner feels on a more equal footing with the scientist.

These criteria are well known to experienced science popularizers, but there was another attraction that seemed counter-intuitive. Journalists were enthralled to discover that there is an equation to describe biscuit dunking. Newspapers published it; TV programmes showed it. More than one radio interviewer even insisted I describe it on air.

All I had done, in fact, was to write down the Washburn equation, derived² in 1921 to describe capillary flow in porous materials:

$$L^2 = \frac{\gamma D t}{4\eta}$$

where t is the time for a liquid of viscosity η and surface tension γ to penetrate a distance L into a fully wettable, porous material whose average pore diameter is D . The equation is strictly true only for capillary flow in a single cylindrical tube in the absence of gravitational effects, but can be extremely accurate for more complex materials, including, as I found experimentally, biscuits. Why this should be so is a very interesting question. In practice, I

could use the Washburn equation to predict how long different biscuits could be safely dunked by the physicist's method, the longest dunkers generally giving the best flavour release (to my palate at least).

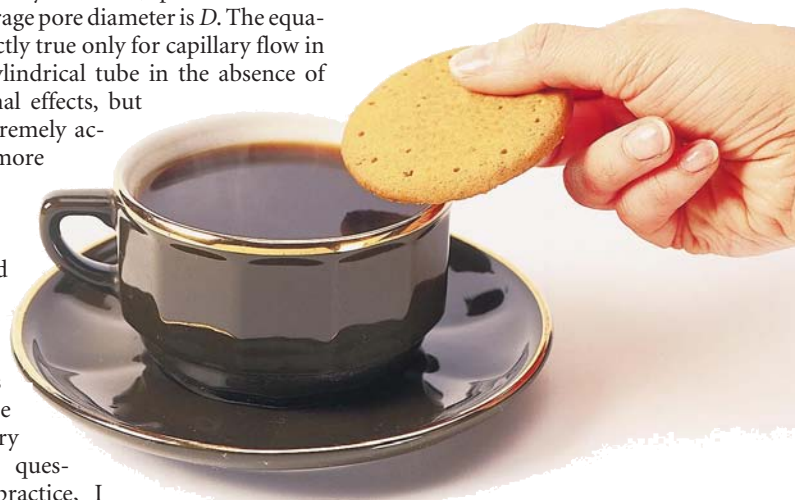
Washburn will be turning in his grave to learn that the media have renamed his work the "Fisher equation". The equation was published in almost every major UK newspaper. The journalists who published it took great care to get it right, some telephoning several times to check. Some even did their own experiments, extending my results. Only one journalist published without checking, provoking the following letter: "Dear Sir, I think there is something wrong with your biscuit dunking equation. Please send me some biscuits for noticing this. Chao Quan (aged 12)."

Such excitement over an equation contradicts the normal publisher's advice to authors — that every additional equation halves the sales of a popular-science book. Why was this so? Let me suggest an answer, relevant to the sharing of more serious science. Scientists are seen by many as the inheritors of the ancient priestly power of the keys, the owners and controllers of seemingly forbidden knowledge. Equations are one key to that knowledge. The excitement of journalists in gaining control of a key was surely a major factor in their sympathetic promotion of the story. By making the Washburn equation accessible, I was able to ensure that journalists unfamiliar with science could use the key to unlock Pandora's box. □

Len Fisher is in the Physics Department, University of Bristol, Bristol BS8 1TL, UK.
e-mail: Len.Fisher@bristol.ac.uk

1. Deegen, R. D. et al. *Nature* **389**, 827–829 (1997).

2. Washburn, E. W. *Phys. Rev.* **17**, 374–375 (1921).



RODNEY FORTE

A transformed view of cyclosporine

Gary J. Nabel

Many transplant patients are given the drug cyclosporine to suppress their immune systems and prevent rejection. But cyclosporine also increases the risk of cancer, always thought to be a side effect of the depressed immune system. A new study shows that cyclosporine directly affects tumour growth and may be the culprit.

Cyclosporine and its relative FK506 are drugs used to treat autoimmune diseases and prevent rejection in bone marrow and organ transplantation. Originally identified by screening for antibiotics from microorganisms, these remarkable drugs suppress the immune system by altering activation of the genes that encode immune factors. However, this suppression must be balanced with the need to maintain enough strength in the immune system to combat infection. So, side effects of cyclosporine treatment are not uncommon, and include an increased risk of cancer in patients who take these drugs long-term.

On page 530 of this issue, Hojo *et al.*¹ describe how they have re-examined the presumed cause of cyclosporine-associated cancers. It was previously held that such cancers result from a failure of the immune system to eliminate cancerous cells — presumably because, when treated with cyclosporine, the immune system cannot detect and respond to proteins associated specifically with the tumours. But Hojo and colleagues suggest a radically different explanation. They find that cyclosporine itself alters the characteristics of several cancerous cell lines *in vitro* and *in vivo*. The authors believe that it does this by inducing the synthesis of another molecule known as transforming growth factor- β (TGF- β).

Hojo *et al.* first found that, when exposed to cyclosporine *in vitro*, cancerous cells become more likely to divide, move and spread — they change shape, show increased mobility and, unlike normal cells, can grow without being anchored to a solid surface. Moreover, when the authors injected different types of tumour cells into immune-deficient mice, more secondary tumours developed in the lungs in the presence than in the absence of cyclosporine. These results challenge assumptions about how cyclosporine-associated tumours arise, and how immune surveillance is involved in the development of cancer. They also raise questions about how cyclosporine suppresses immune function, suggesting a more general role than previously thought for cyclosporine-dependent signalling pathways in human disease.

The mechanism of cyclosporine action

has been described by several laboratories (reviewed in refs 2, 3). It was thought to block the immune system by inhibiting signalling through the T-cell receptor. The effect of this is to prevent the production of cytokines that would normally stimulate an immune response (Fig. 1). But this explanation is not necessarily complete. First, when one of these cytokines, interleukin-2, is disrupted in transgenic animals, the effect on immune function is not the same as treatment with cyclosporine^{4–7}. Second, if another molecule in the proposed signalling pathway (the nuclear factor of activated T-cells, NF-AT) is knocked out, again the effects do not match treatment with cyclosporine⁸.

An alternative explanation comes from the fact that cyclosporine stimulates the production of TGF- β ^{9,10}. We do not know how it does this, but it is an important avenue of further investigation because several previously mysterious side effects of cyclosporine can be explained by the induction of TGF- β (Fig. 2). For example, cyclosporine can cause liver damage, and thickening and scarring of the kidneys and skin — an effect also seen with increased TGF- β ¹¹. Because TGF- β is itself a well-known and potent immunosup-

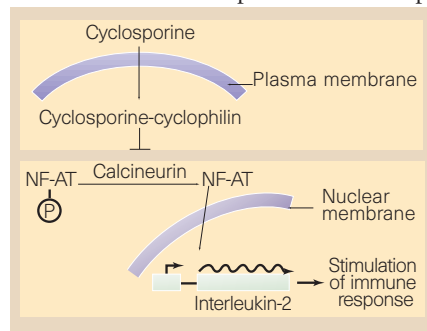


Figure 1 Previously defined mechanism for the action of cyclosporine. After binding to cyclophilin, cyclosporine inhibits calcineurin, which normally removes a phosphate group from members of the nuclear factor of activated T-cells (NF-AT) family. This dephosphorylation normally promotes movement of NF-AT to the nucleus, where it stimulates expression of the genes for interleukin-2 and other stimulatory cytokines. Inhibition by cyclosporine blocks cytokine expression and prevents an immune response.

pressive agent, increased production of TGF- β — as well as the inhibition of cytokine production — could explain the activity of cyclosporine.

Hojo *et al.*¹ have now recognized that, as well as affecting immune function, TGF- β might alter the behaviour of cancerous cells. The remarkable changes that they observed in cyclosporine-treated cancerous cell lines were reversed when they added an antibody that bound to (and therefore blocked the action of) TGF- β . In the same way, the increased spread of cancer cells *in vivo* was also blocked by the TGF- β antibody. These results indicate that treatment with cyclosporine does not increase tumour growth indirectly, by a failure of the immune system, but rather directly through a non-immune mechanism that acts on the tumour itself via TGF- β receptors.

Hojo and colleagues' study is certainly provocative, calling into question the proposed mechanism by which cyclosporine induces secondary cancers. But several caveats should be noted. For example, we do not know whether cyclosporine has a similar effect on precancerous cells, or whether it is involved in converting cells from a benign to a cancerous state. Nonetheless, the data do indicate that cyclosporine and its relatives could exacerbate tumour growth in patients with existing tumours.

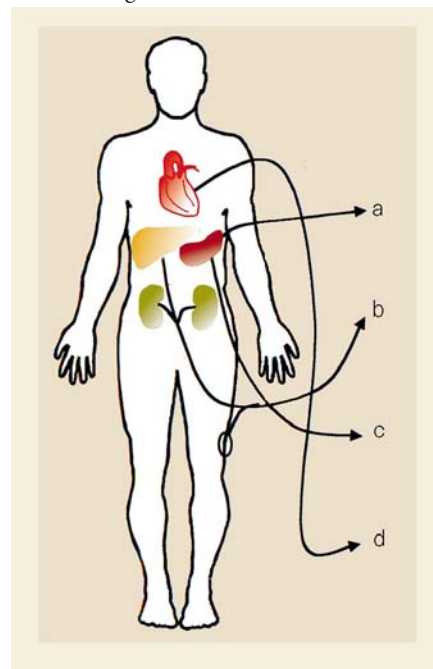


Figure 2 Emerging ideas about the action of cyclosporine. Hojo and colleagues¹ have shown that cyclosporine may cause cancer by stimulating production of the transforming growth factor- β (TGF- β), although it is not yet known how it does this. TGF- β may then: a, stimulate the growth of existing cancers; b, cause liver damage, and thickening and scarring of the skin or kidneys; or c, suppress immune responses. d, Cyclosporine may also be used to treat heart disease by inhibiting thickening of the heart muscle.

Other studies point to a broader involvement of the signalling pathway triggered by cyclosporine in tissues outside the immune system. Molkentin, Olson and colleagues¹², for example, have reported that increasing the expression of activated NF-AT3 in the heart muscle of transgenic mice leads to enlargement of the cardiac tissue (cardiac hypertrophy). Cardiac hypertrophy can be inhibited by giving mice cyclosporine¹³, although it is not clear whether TGF- β is the mediator involved here. An alternative explanation may come from the finding that cyclosporine-mediated inhibition of calcineurin blocks the expression of muscle-specific genes, preventing hypertrophy¹⁴. Whatever the mechanism, these results indicate that a cyclosporine-like agent could be used to treat heart disease. However, such treatment would probably not be given unless the effects of cyclosporine on the immune system could somehow be reduced.

Hojo and colleagues' study will raise concerns about the increased incidence of cancer associated with cyclosporine treatment. This complication is, however, a well-known side effect of cyclosporine therapy, one that is currently balanced against the need to treat life-threatening diseases. The new observations do not alter this risk, nor do they sug-

gest that any additional precautions be taken beyond those already recognized. But they do provide an insight into how these cancers come about, and may be useful in treating them. The results could also provide a lead in the search for new immunosuppressive drugs that might be more selective. But cyclosporine, whose mechanism of action was thought to be well understood, remains enigmatic in many respects. □

Gary J. Nabel is at the University of Michigan, 1150 West Medical Center Drive, 4520 MSRB I, Ann Arbor, Michigan 48109-0650, USA.
e-mail: gnabel@umich.edu

1. Hojo, M. et al. *Nature* **397**, 530–534 (1999).
2. Schreiber, S. L. & Crabtree, G. R. *Immunol. Today* **13**, 136–142 (1992).
3. Rao, A. J. *Leukocyte Biol.* **57**, 536–542 (1995).
4. Schorle, H., Holtschke, T., Hunig, T., Schimpl, A. & Horak, I. *Nature* **352**, 621–624 (1991).
5. Kundig, T. M. et al. *Science* **262**, 1059–1061 (1993).
6. Horak, I., Lohler, J., Ma, A. & Smith, K. A. *Immunol. Rev.* **148**, 35–44 (1995).
7. Contractor, N. V. et al. *J. Immunol.* **160**, 385–394 (1998).
8. Xanthoudakis, S. et al. *Science* **272**, 892–895 (1996).
9. Li, B. et al. *J. Exp. Med.* **174**, 1259–1262 (1991).
10. Khanna, A., Stenzel, K. H. & Suthanthiran, M. *Transplantation* **57**, 577–582 (1994).
11. Border, W. A. & Noble, N. A. *N. Engl. J. Med.* **331**, 1286–1292 (1994).
12. Molkentin, J. D. et al. *Cell* **93**, 215–228 (1998).
13. Sussman, M. A. et al. *Science* **281**, 1690–1693 (1998).
14. Chin, E. R. et al. *Genes Dev.* **12**, 2499–2509 (1998).

Particle physics

The undemocratic proton

David J. Miller

Since the proton is the lightest stable composite object in the Universe we might expect it to be the simplest. Yet its structure is much more complicated than we can explain with current theories, as reported in *Physical Review Letters* by the HERMES collaboration¹ at the HERA accelerator in Hamburg and by the NuSea collaboration² at the Fermilab Tevatron near Chicago. HERMES

used semi-inclusive deep-inelastic scattering of a positron beam to identify the contribution of individual quarks to the momentum of the proton. Both groups see an excess of virtual down-antiquarks over up-antiquarks, contradicting predictions that the two flavours of antiquark should be present inside the proton in equal numbers.

Elastic scattering is what happens with

billiard balls — they collide and bounce without changing their identity and with no significant loss of kinetic energy. Inelastic scattering usually means that one or both of the colliding objects has been changed by the collision, in that they may be deformed or broken into pieces like a billiard ball shattered by a bullet. Deep-inelastic scattering is a subtle amalgam of the two, as if a bullet shatters a billiard ball but in the process makes a quasi-elastic collision with a heavy, compact ball-bearing buried inside the target. Particle physicists began to believe in the reality of quarks in the early 1970s (ref. 3), when deep-inelastic scattering experiments with electrons (bullets) on protons (billiard balls) gave far more large-angle electron scatters than expected. The results can be explained only if there are heavy, compact objects (the quarks, equivalent to the buried ball-bearings) inside the protons.

This raises an obvious question: if an electron has a quasi-elastic scatter on a quark, should we not see the quark recoiling at a large angle too, leaving the shards of the proton behind? The answer is “Yes, but it is more complicated than that”. What obscures the simple picture is a property of quarks that makes them harder to study than most other elementary particles, the fact that a single free quark can never be seen on its own. They are like the poles of a bar magnet — cut it in two and you don't get two free poles, you get two smaller magnets each with two poles. According to the successful theory of quantum chromodynamics (QCD), quarks are always surrounded by a local cloud of force-carrying particles called gluons, which act like the magnetic field around a magnetic pole. If we scatter an electron violently from a quark in a proton, then the quark begins to recoil like a free particle. But when it gets more than about 1 femtometre (10^{-15} m) from the remains of the proton, it stretches the gluon field sufficiently to allow a new quark–antiquark pair to be produced, and the antiquark sticks to the original quark and makes a meson.

In highly energetic collisions, many quark–antiquark pairs are created, producing a jet (Fig. 1) of mesons (see ref. 4 for a qualitative explanation, or ref. 5 for the full treatment). But QCD theory says that the meson with the largest momentum will contain the original recoiling quark. A π^+ meson contains an up-quark and a down-antiquark; a π^- contains a down-quark and an up-antiquark, so by looking at the fastest meson we can infer something about which kind of quark (or antiquark) in the target proton was hit. Experiments that pick out this leading meson for special attention are making use of semi-inclusive deep-inelastic scattering.

That is what the HERMES collaboration¹ has done: they scattered 27.5 GeV positrons (the same as electrons for these purposes) from both protons and neutrons contained

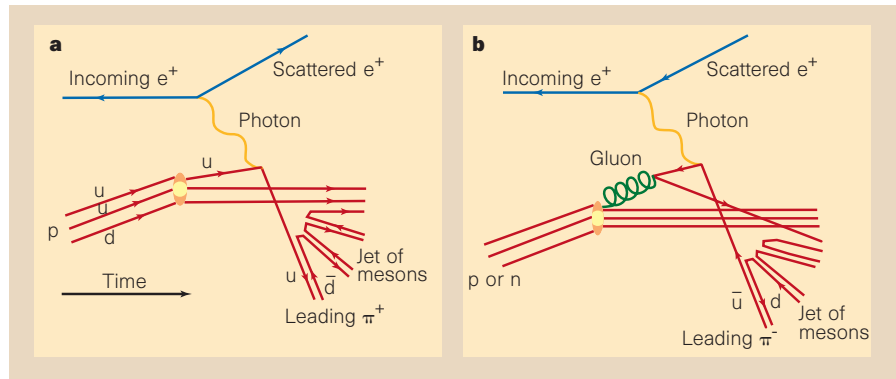


Figure 1 Principle of the HERMES experiment¹. a, Feynman graph for semi-inclusive deep-inelastic positron scattering from a valence up-quark in a proton. The exchanged photon is the carrier of the electromagnetic force. If the leading meson is charged, it has to be a π^+ because the recoiling quark is an up-quark. Arrows pointing backwards in time represent antiparticles. b, Semi-inclusive deep-inelastic scattering from a virtual up-antiquark ($\bar{u}$) in the sea, formed by pair production from a virtual gluon. The leading charged meson in this case has to be a π^- , whereas if the scattering were from a down-antiquark it would be a π^+ .

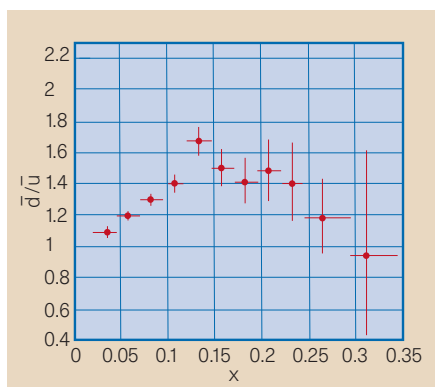


Figure 2 The excess of down-antiquarks in the proton sea. The ratio of $\bar{d}/\bar{u}$ as a function of the fraction x of the momentum of the target proton carried by the struck quark, from the NuSea experiment².

in puffs of hydrogen, deuterium or helium gas, which were injected into the vacuum chamber of the accelerator. As well as detecting the recoiling electron they observed the fastest meson coming out on the other side, and recorded whether it was a π^+ or a π^- . The velocity of the recoiling electron tells them the fraction x of the momentum of the proton or neutron that was carried by the struck quark. When x is large (more than about 0.2) we expect the struck quark to be a valence quark. The up and down valence quarks give particles their charge. There are two up- and one down-quarks in a proton, and two down- and one up-quarks in a neutron. Figure 1a shows the Feynman diagram for scattering off a valence up-quark in a proton, with the intermediate photon carrying momentum from the incident positron to the struck up-quark. At small x we also expect extra scatters, from the sea of short-lived, virtual quark-antiquark pairs produced by quantum fluctuations in the gluon field (Fig. 1b).

Simple QCD theory had predicted that this sea should contain equal numbers of up- and down-antiquarks, but the HERMES results show that there is a clear excess of down-antiquarks over the number of up-antiquarks, and the effect becomes clearer for smaller values of x . The NuSea experiment² achieved a similar result by a different method, which involved production of muon pairs in proton scattering from hydrogen and deuterium targets (Fig. 2).

Simple QCD theory does not predict an 'asymmetric sea' inside the proton; gluons should be democratic in the kinds of quark-antiquark pairs they generate. So far, the only satisfactory explanation for the excess of down- over up-antiquarks comes from more old-fashioned ideas, going back to the picture of the strong nuclear interaction that we had before QCD. Several theorists have shown that if the proton and neutron are assumed to contain virtual pions — as well as virtual quarks, antiquarks and

gluons — then the asymmetry can arise quite naturally (see references in ref. 2). This does not mean that QCD is wrong, just that we still don't know how to use it to calculate all of the subtle properties of such a complicated object as a proton. □

David J. Miller is in the Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK.

e-mail: djm@hep.ucl.ac.uk

1. Ackerstaff, K. *et al.* (HERMES collaboration) *Phys. Rev. Lett.* **81**, 5519–5523 (1998).
2. Hawker, E. A. *et al.* (NuSea collaboration) *Phys. Rev. Lett.* **80**, 3715–3718 (1998).
3. Miller, G. *et al.* (SLAC-MIT collaboration) *Phys. Rev. D* **5**, 528–544 (1972).
4. Miller, D. J. *Nature* **349**, 379–387 (1991).
5. Ellis, R. K., Stirling, W. J. & Webber, B. R. *QCD and Collider Physics* (Cambridge Univ. Press, 1996).

Colour vision

A patchwork of cones

Heinz Wässle

If you look closely at the screen of a colour television, you'll see a precise mosaic of red, green and blue pixels. At a normal viewing distance, however, our eyes cannot resolve these individual pixels and, depending on their relative intensities, we perceive thousands of different colours. Television screens (and now PC monitors) are the most popular applications of the Young-Helmholtz theory of trichromacy¹, which states that human colour vision is based on three types of cone in the retina. Yet over 100 years after Helmholtz formulated his ideas, we have a paradox. Although engineers can construct perfect colour monitors, neurobiologists still do not know the details of the circuit that subserves trichromatic colour vision in humans, despite contributions from psychophysics, histology, electrophysiology, spectroscopy and molecular genetics. Now, on page 520 of this issue, Roorda and Williams² go some way towards addressing this problem.

Why do we not know more about human trichromacy? For one, there is no other species of animal whose retina could serve as a model. Birds, reptiles and fishes have elaborate colour vision, but their retinas differ greatly from those in primates. Many mammals have an evolutionarily ancient, dichromatic form of colour vision, but there are no known mammalian trichromates besides the

primates (monkeys, apes and man)³.

The primate retina has three types of cone, each containing a different photopigment (opsin)⁴. Their peak sensitivities lie in the violet (short wavelength S-cones), green (medium wavelength M-cones) and yellow-green (long wavelength L-cones) regions of the colour spectrum⁵. So far, only S-cones have been distinguished by microscopy and, on this basis, their circuitry has been studied in detail. As in other mammalian retinas, the S-cones make up 5–10% of all cones and form a regular mosaic. Signals from the S-cones are transmitted by special types of bipolar and ganglion cells^{6,7}, and there is also evidence⁸ that the S-cone signal takes a separate route through the thalamus to the visual cortex.

Unfortunately L- and M-cones cannot be distinguished by their shapes or other anatomical means — their opsins differ in only 15 out of 363 amino acids, so nobody has yet succeeded in producing specific antibodies to them. But Mollon and Bowmaker⁹ showed one way round this difficulty. By measuring the absorption curves of all cones in a patch of monkey retina *in vitro*, they defined the distribution of the L- and M-cones. They found equal numbers of the two cone types, randomly distributed.

Roorda and Williams² now take this analysis of the cone mosaic much further

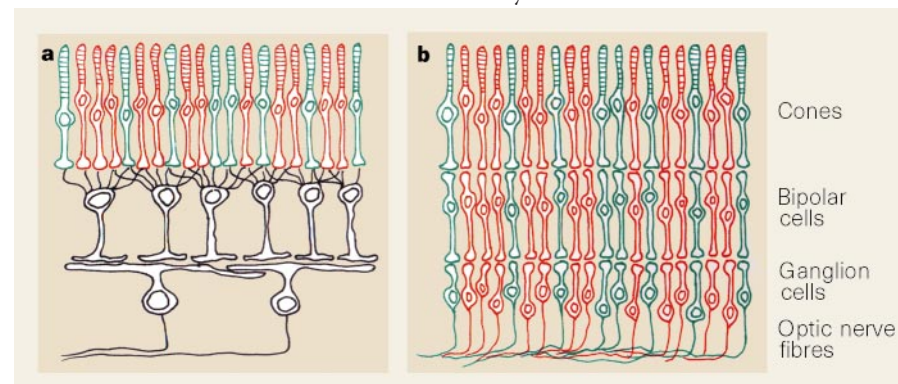


Figure 1 Organization of the retina in different animals. a, In most mammalian retinas, bipolar cells pool the signals of several neighbouring cones and, in turn, ganglion cells pool the signals of several bipolar cells. b, In the central retina of primates, each cone is connected to a midget bipolar cell and from this to a midget ganglion cell which inherits the chromatic signal of that cone. Roorda and Williams² have used a new technique to show how the different types of cone are distributed in the human eye.

and, for the first time, show the distribution of L-, M- and S-cones in the living human eye. They used an elaborate optical system ('adaptive optics') that corrects for the bad imaging quality of the cornea and lens. By differentially bleaching the retina with strong blue or red light, the authors identified L-, M- and S-cones in an extended area of retina. The proportion of L- to M-cones was nearly 4:1 in one male subject, yet closer to 1:1 in a second subject — each with normal colour vision when tested psychophysically. The L- and M-cones were randomly distributed in the two subjects, and both had large patches in which either L- or M-cones were missing.

There is no doubt that Roorda and Williams' technique, combined with psychophysical performance measurements in the same subject, will present new insights into colour vision and its variations among individuals. The patchy distribution of M- and L-cones implies that the spatial grain for colour is coarser than that for intensity, explaining why our eyes fuse the coloured pixels of the television monitor. Now we can apply this technique of imaging L- and M-cones to the monkey eye, allowing us to work out the circuitry that subserves trichromacy.

Why has trichromacy evolved in primates and not in other mammals? In most mammals, bipolar cells pool the signals of several neighbouring cones, then ganglion cells pool the signals of several converging bipolar cells (Fig. 1a). In such a highly convergent system, any chromatic information introduced into the cone mosaic by, say, a mutation that creates L- and M-cones, would be lost within the retina and would never reach the brain¹⁰.

But the situation was different 30 million years ago for primates. During evolution, the primate eye and retina have been optimized for the highest spatial resolution. This required a high density of cones and a low ratio of cones to ganglion cells in the 'acuity pathway'. The anatomical limits for this optimization were reached when each cone was connected, through a midjet bipolar cell, to a midjet ganglion cell (Fig. 1b). In this way, a 'private line' to the brain was established. Only after evolution of this one-to-one connection in the central retina did a subsequent mutation create the patchy distributions of L- and M-cones at random locations¹¹. The one-to-one midjet-cell system of the central retina could then transmit this chromatic information to the brain. Gradually, the selective advantage of trichromatic vision must have led the colour-processing pathways to proliferate in cortical, and perhaps even subcortical, centres. The well-known plasticity of the brain could easily have allowed such changes. The idea that trichromacy piggybacks on the high-acuity system also indicates that the midjet ganglion cells do a double duty in visual signalling — an idea that has been promoted for some years¹². □

Heinz Wässle is at the Max Planck Institut für Hirnforschung, Deutschordenstrasse 46, D-60528 Frankfurt/Main, Germany.
e-mail: waessle@mpih-frankfurt.mpg.de

1. von Helmholtz, H. *Treatise on Physiological Optics* (ed. Southall, J. P.) 3 Vols (Dover, New York, 1962).
2. Roorda, A. & Williams, D. R. *Nature* **397**, 520–522 (1999).
3. Jacobs, G. H. *Biol. Rev.* **68**, 413–471 (1993).
4. Nathans, J., Merbs, S. L., Sung, C. H., Weitz, C. J. & Wang, J. *Annu. Rev. Genet.* **26**, 403–424 (1992).
5. Schnapf, J. L., Kraft, T. W. & Baylor, D. A. *Nature* **325**, 439–441

(1987).

6. Kouyama, N. & Marshak, D. W. *J. Neurosci.* **12**, 1233–1252 (1992).
7. Dacey, D. M. & Lee, B. B. *Nature* **367**, 731–735 (1994).
8. Martin, P. R., White, A. J. R., Goodchild, A. K., Wilder, H. D. & Sefton, A. E. *Eur. J. Neurosci.* **9**, 1536–1541 (1997).
9. Mollon, J. D. & Bowmaker, J. K. *Nature* **360**, 677–679 (1992).
10. Wässle, H. & Boycott, B. B. *Physiol. Rev.* **71**, 447–480 (1991).
11. Mollon, J. D. *J. Exp. Biol.* **146**, 21–38 (1989).
12. Ingling, C. R. Jr & Martinez-Urieag, E. *Vision Res.* **23**, 1495–1500 (1983).

Oceanography

Bacteria and silica cycling

Victor Smetacek

As a graduate student who enjoyed counting plankton under the microscope, I felt frustrated when confronted with an occasional sample that contained almost nothing except a few bacteria flitting or gliding about on the slide. The bacteria were apparently responsible for the barrenness of the slide, and had thrived because the sample had escaped its shot of preservative. Mysteriously, the silica shells of the diatoms had also vanished, although they were beautifully preserved in the treated samples. The observation contradicted the view that diatom shells are of no nutritional value and that their dissolution is controlled by physico-chemical and not biological factors.

This dogma has now itself been dissolved (and hence my hunch borne out) by the elegant study by Bidle and Azam on page 508 of this issue¹. They show that bacteria do indeed 'gnaw' enzymatically into diatom shells, thereby hastening their dissolution rates significantly (Fig. 1). The authors also show that adding proteases (enzymes that break down proteins) had the same effect as live bacteria, indicating that a protein coating provides protection from dissolution.

Diatoms dominate phytoplankton blooms in the ocean, so the finding adds to our growing understanding of how planet-

ary elemental cycles are geared to each other. In the emerging view, wind-blown dust from the continents increases ocean productivity by providing iron², which is often a limiting element. In large areas of the open ocean — the high-nutrient, low-chlorophyll (HNLC) regions — iron deficiency constrains growth rates not only of phytoplankton² but also of bacteria³. In some regions, silicic acid (dissolved silica) is in short supply relative to the major nutrients nitrate and phosphate⁴. For instance the N:P ratio across the Southern Ocean varies by less than 10% but the Si:N ratio ranges from 0.15 in the north to 3 in the south. Because diatom demand for Si:N is about 1, the discrepancy in dissolved nutrient ratios indicates that nitrogen is recycled more efficiently within the surface layer than silica. Hence the finding that microbial breakdown of protein also enhances silica dissolution¹ has interesting implications, because until now diatoms have been thought to protect their surface by secreting mucus (carbohydrates).

The world ocean is strongly undersaturated in silicic acid but its cycle is in steady state, with the annual input from rivers — about six tera (10^{12}) moles — being balanced by burial of diatom shells in sediments⁵. However, the sediments underlying the productive ocean

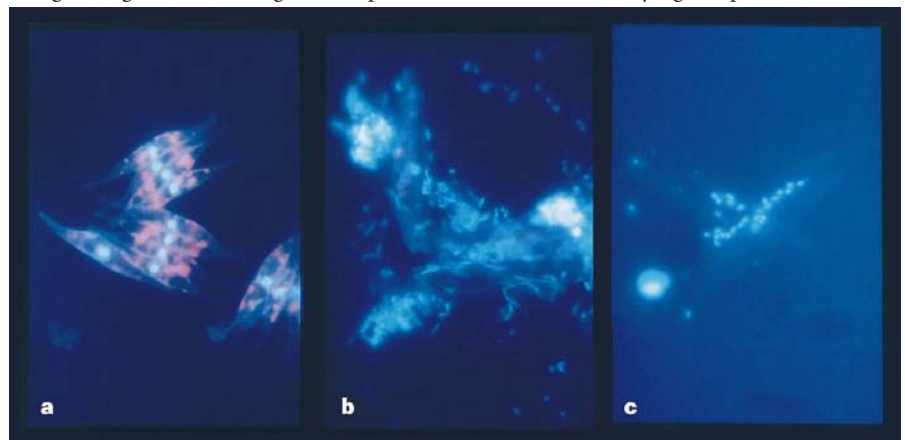


Figure 1 The case of the disappearing diatoms — an example of Bidle and Azam's experiments¹. These pictures are a time series, running left to right, showing bacterial attack on the silica shells of the diatom *Cylindrotheca fusiformis*. In a, time zero, the diatoms' integrity is intact. b and c show their progressive dissolution after two and seven days, the bacteria staining bright blue. (Courtesy F. Azam.)

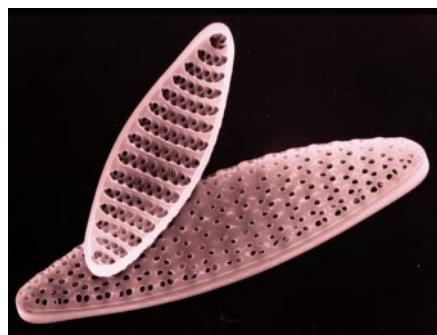


Figure 2 Fortress *Fragilariopsis kerguelensis*, the shells of which dominate the diatom ooze under the Antarctic Circumpolar Current. The valves are about 30 μm long and 10 μm wide; shown here are the inner (top) and outer surfaces of the valves, which are the main component of the silica cell wall. (Courtesy F. Hinz and R. M. Crawford, Alfred Wegener Institute.)

margins and the North Atlantic, where most of the diatom blooms occur, accumulate only minor amounts of silica. Instead, the bulk of it is buried under the iron-limited HNLC areas, about as far from riverine sources as one can get. With the information now available, the discrepancy can be explained: iron-limited diatoms make thicker silica shells^{6,7}, and breakdown of protective protein is slowed because bacterial growth is iron-limited. So a larger percentage of the shells in HNLC regions survive and are buried.

However, the ooze accumulating on the ocean floor is composed of relatively few species: the shells of most diatoms, including some shown to thicken under iron deficiency, are absent. Indeed, about half the diatom ooze accumulating under the Antarctic Circumpolar Current, which constitutes fully two-thirds of the entire ocean burial, consists of the shells of one species, *Fragilariopsis kerguelensis*⁸ (Fig. 2). Apparently there are large species-specific differences in the susceptibility to dissolution of the shells. This is also confirmed by the findings of Bidle and Azam¹, who attribute them to differences in architecture of the silica structure. So shell thickness is not the only determinant of preservation in sediment. That, however, leaves the question as to the advantage, if any, of making different shells that dissolve at different rates after death of the cell.

Most ocean phytoplankton is eaten almost as fast as it grows. Diatom growth rates are in the same range as those of smaller algae, yet, because diatoms accumulate to form blooms, their survival rate must be much longer. Indeed this is a prerequisite for the diatom strategy of storing nutrients in a vacuole for subsequent cell divisions. Because an adequate amount of silicic acid is an absolute requirement for all diatoms, the silicified cell walls must play a central role in their success story. Clearly the larger size, tough walls, spines and chitin threads of the diatoms deter most of the protozoan grazers

that consume the bulk of the smaller phytoplankton. The crustacean zooplankton, armed with mandibles and gizzards, are better equipped to deal with diatoms, and they leave a trail of broken shells in their wake. Bidle and Azam now show how this siliceous debris can be recycled back to the diatoms by bacteria, enabling surviving cells to divide and sustain the population presence in the surface layer. Such recycled silica must maintain the diatoms of the regenerating communities characteristic of the nitrogen-limited subtropical gyres.

If silica shells indeed represent the first line of defence in the diatoms, then strengthening them will increase the prospects for survival. Indeed, the shells of *Fragilariopsis* look more like fortresses than water tanks (to support the vacuole, as is widely believed). Such species will have invested in the construction of more durable shells that are more likely to end up in the sediment, keeping the ocean undersaturated in silicic acid.

The biochemistry of silica (we also have

some in our hair and nails) is still poorly known, and proteins with unique structures and properties are implicated in diatom shell formation⁹. Perhaps it is these proteins that remain to shield the silica from dissolution. But we are left with an enigma: how do diatoms prevent bacteria from attacking the protein on the outside of their shells? Can the pharmaceutical or anti-fouling paint industries learn something or even profit from them? □

Victor Smetacek is at the Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, D-27570 Bremerhaven, Germany.
e-mail: vsmetacek@awi-bremerhaven.de

1. Bidle, K. & Azam, F. *Nature* **397**, 508–512 (1999).
2. Falkowski, P. G., Barber, R. T. & Smetacek, V. *Science* **281**, 200–206 (1998).
3. Pakulski, J. D. *et al.* *Nature* **383**, 133–134 (1996).
4. Dugdale, R. C., Wilkerson, F. P. & Minas, H. J. *Deep-Sea Res.* **42**, 697–719 (1995).
5. Treguer, P. *et al.* *Science* **268**, 375–379 (1995).
6. Hutchins, D. A. & Bruland, K. W. *Nature* **393**, 561–564 (1998).
7. Takeda, S. *Nature* **393**, 774–777 (1998).
8. Verity, P. G. & Smetacek, V. *Mar. Ecol. Prog. Ser.* **130**, 277–293 (1996).
9. Kröger, N. & Sumper, M. *Protist* **149**, 213–219 (1998).

Active galactic nuclei

Unstable by design

Aneta Siemiginowska and Martin Elvis

The Wright brothers made their first aircraft manoeuvrable by making it unstable. Now it seems that quasars — the most luminous type of active galactic nuclei — also have to be unstable to make them work. This finding may resolve a long-standing problem with the quasars' fuel supply and allow studies into the distribution of massive black holes in the Universe.

Flattened disks are common in astronomy (recall Saturn's rings and newly forming planetary systems). When a disk forms deep in a gravitational potential well, it can become one of the most efficient radiation engines possible, and so can power the most prolific of radiation sources, the quasar. But if quasars were continuously radiating, they would grow too large and run out of fuel. On the other hand, if quasar disks were unstable, then quasars would be intermittent¹, and mostly non-radiating.

In the *Astrophysical Journal*, Burderi *et al.*² summarize a decade of work on quasar disk instabilities^{3–7}, and show that even if the disk is irradiated from above it will be unstable. Quasar disks are usually illuminated by external radiation (such as coronae, jets or radiation scattered off nearby gas clouds), which may prevent the occurrence of the ionization instability. Burderi *et al.* include disk irradiation in their model and apply a global instability criterion to the outer edge of the accretion disk. They show that the ionization instability should be present in quasar accretion disks. With this difficulty solved, theorists can now tackle some previously unad-

dressable puzzles, such as the mass distribution of massive black holes and their evolution with cosmic time.

How do disks become efficient enough to power quasars and other active galactic

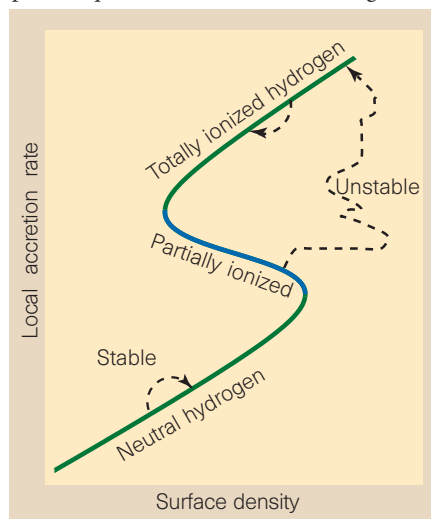


Figure 1 Relationship between local surface density and accretion rate in active galactic nuclei. Any perturbation in local accretion rate (dashed lines) on the neutral or totally ionized branches of the curve (green) can be balanced by a change in surface density and the disk remains stable. But if this perturbation happens when the disk is on the partially ionized branch (blue), then the disk is unstable and the only possible outcomes are for the disk to become totally ionized, or neutral. The study by Burderi *et al.*² concludes that most, if not all, observed active galactic nuclei are unstable.

nuclei? If matter falling (accreting) onto a compact object, such as a black hole, has enough angular momentum, then it forms a rotating accretion disk. Friction between adjacent disk orbits transports excess angular momentum outward, allowing the matter to flow inward. Heat energy generated from this viscosity (mechanical or magnetic, it is not clear which) is radiated at the disk surface. The amount of power released by the disk depends on how far down the disk extends into the gravitational well of the central object. The most extreme case is if the central object is a rotating black hole, when as much as 40% of the rest-mass energy can be emitted.

Accretion onto a supermassive black hole (of mass $\sim 10^8$ solar masses) was accepted as the most efficient way to power a quasar 30 years ago⁸. Stationary accretion disk models have been widely used to describe the optical and ultraviolet emission spectra of quasars. But quasar luminosities require a large amount of accreting matter (about 2 solar masses per year), which cannot be sustained throughout the quasars' entire lifetime, or we would find massive remnants of dead quasars today in nearby galaxies, which we don't⁹. This fuelling problem could be solved if the intermittent or one-shot 'quasar events' in the life of a galaxy are triggered by unstable disks.

Over the past decade, theorists^{6,7} have gone beyond the physics of steady quasar disks by considering the 'hydrogen ionization' instability (Fig. 1). This is the same instability that puffs up, and causes oscillations in, red giant stars. In some binary star systems¹⁰ it causes huge outbursts, with an increase in luminosity by a factor of about 10^4 in just a few days. The instability is caused by a temporary increase in the opacity of the disk material during hydrogen ionization. The disk is stable when hydrogen is neutral because the viscous heating rate is balanced by the cooling rate. In the act of ionizing hydrogen, ultraviolet photons are absorbed, and the cooling rate suddenly becomes too small to compensate for the heating rate, puffing up the disk locally. Balance is restored when the hydrogen becomes totally ionized, and so transparent in the ultraviolet.

It was not clear that this instability applied to quasars³, because if it did it would operate over 10^4 – 10^6 years, too long for mortals to study directly. Burderi *et al.*² studied the general conditions for the instability to occur in quasars and show that even if the disk is externally irradiated, which appears likely, the instability will still develop. They conclude that all observed active galactic nuclei are in the outburst phase.

We have calculated¹¹ that the outburst phase lasts for about 10% of the entire life of a quasar accretion disk. This means that about 10% of galaxies at a given time should exhibit quasar activity (Fig. 2). The corresponding long periods of quiescence resolve the old

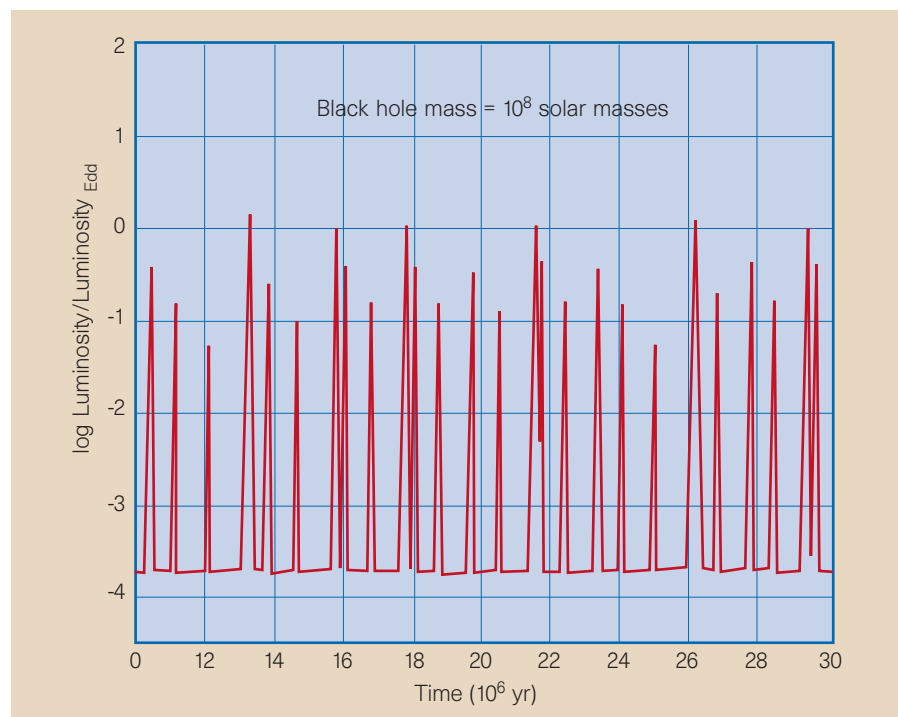


Figure 2 The time history (light curve) of the output of a quasar accretion disk, with the ionization instability included, shows large luminosity outbursts (from ref. 11). Luminosity is expressed in the Eddington luminosity. It shows that most ($\sim 75\%$) sources will be quiescent at a given moment, whereas only $\sim 10\%$ of sources would be in outburst phase. The remaining $\sim 15\%$ of sources would be in a transition state with moderate luminosities and some activity. The conclusion that there must be a large population of 'dormant' quasars is supported by a number of observations¹².

fuelling problem of quasars. If the disk is unstable, then a steady mass inflow into the disk around a black hole of 10^8 solar masses is only about 0.2 solar masses per year. The black hole then grows by only a factor of two or so, even after a Hubble time (this is the age of the expanding Universe $= 10^{10} h^{-1}$ years, where $h = 0.01 H_0 \text{ km s}^{-1} \text{ per megaparsec}$ as defined by the Hubble constant, H_0). During quiescence the disk emission is 10^5 times fainter, and is masked by the emission of the galaxy in which the quasar lies, so many galaxies should contain dormant quasars¹².

During disk outbursts, radiation from the disk reaches the highest possible luminosities, restricted only by the Eddington limit, which is defined by the balance between the total gravitational force and radiation pressure. If radiation pressure generated within an object exceeds the gravitational force, then the outer layers of the object can be blown away. For stars this limit gives a maximum stable mass. The accretion flow during an outburst is extremely rapid, releasing the matter stored up during quiescence at a rate of about two solar masses per year. Rapid outflows or jets of nuclear material often occur during these outbursts. Studies of the fast-moving 'superluminal knots' in relativistic jets suggest intermittent ejection, with timescales comparable to the ionization instability¹³. In the past two years a few binary stars have been found to eject 'superluminal blobs' of material, which can

be seen at radio wavelengths, and this ejection follows a change in the accretion flow, as measured by X-ray emission¹⁴. Could this be a small-scale version of radio galaxy behaviour, when the central black hole mass is of the order of a few solar masses?

Now we believe that the ionization instability does operate in quasars, we can develop physics-based models of quasars. Although individual quasars change too slowly to observe, a population of quasars would contain predictable numbers of sources in different states of activity. The distribution of the luminosity of quasars (their 'luminosity function') is recorded both locally and at great distances (high redshifts). The instability can be used to predict the distribution of luminosities for a given black hole mass. Then we can use the luminosity function to find the mass distribution of massive black holes, and, by comparing luminosity functions at different redshifts, their evolution with cosmic time. We have already tried this with some success, although the model was quite simple¹¹. There is a new maturity to the theory of quasar accretion disks, and this is bringing questions about the black hole population of the Universe out of the realm of speculation, and into physics. □

Aneta Siemiginowska and Martin Elvis are in the Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA.

e-mails: aneta@head-cfa.harvard.edu



100 YEARS AGO

A series of tables, showing the differences between Greenwich mean time and the civil times used in various parts of the world, compiled by Prof. John Milne, F.R.S., is published in the February number of the *Geographical Journal*. The names of places in the tables are arranged in alphabetical order, and the amount by which the time used at each is fast or slow of Greenwich mean time is indicated. ... It is pointed out that the Chinese at most places use an approximate apparent solar time, obtained from sun-dials. At Tientsin the civil time is determined by the municipal chronometer, which, however, has sometimes been known to have an error of three minutes. The Persians keep sun time, watches being set at sunset. In Teheran there is a midday gun fired by the time shown on a sun-dial. But a few minutes makes no difference in Persia; the railway trains start when full or when required, and Persian telegraphists do not give time of issue or receipt of telegrams.

From *Nature* 9 February 1899.

50 YEARS AGO

The nucleic acids and their derivatives are fundamental constituents of biological systems; but until recently workers have lacked precise micro methods for their study. Vischer and Chargaff, and recently Hotchkiss and also Reichard, have shown that partition chromatography may be used to separate and characterize certain derivatives. The methods employed by these workers to detect the spots on the chromatograms were, however, either limited in their scope or necessitated the analysis of a large number of fractions selected arbitrarily from the columns. We have approached this problem in a different way. Purine and pyrimidine derivatives are characterized by an intense ultra-violet light absorption in the region of 2650 Å., and they can be detected in amounts of 10 µgm. or less by a simple photographic technique.— Roy Markham, John D. Smith

From *Nature* 12 February 1949.

Many more extracts like these can be found in *A Bedside Nature: Genius and Eccentricity in Science, 1869–1953*, a 266-page book edited by Walter Gratzer. Contact Lisa O'Rourke. e-mail: l.orourke@nature.com

elvis@head-cfa.harvard.edu

1. Shields, G. A. & Wheeler, J. C. *Astrophys. J.* **222**, 667–674 (1978).
2. Burderi, L., King, A. R. & Szuksiewicz, E. *Astrophys. J.* **509**, 85–92 (1998).
3. Lin, D. N. C. & Shields, G. A. *Astrophys. J.* **305**, 28–34 (1986).
4. Clarke, C. J. *Mon. Not. R. Astron. Soc.* **235**, 881–894 (1989).
5. Cannizzo, J. K. *Astrophys. J.* **385**, 94–107 (1992).
6. Mineshige, S. & Shields, G. A. *Astrophys. J.* **351**, 47–61 (1990).
7. Siemiginowska, A., Czerny, B. & Kostyunin, V. *Astrophys. J.* **458**, 491–507 (1996).

8. Lynden-Bell, D. *Nature* **223**, 690–691 (1969).

9. Cavaliere, A. & Padovani, P. *Astrophys. J.* **333**, L33 (1988).
10. Smak, J. I. *Acta Astron.* **32**, 199–211 (1982).
11. Siemiginowska, A. & Elvis, M. *Astrophys. J.* **482**, L9 (1997).
12. Ho, L. C. in *The Physics of Liners in View of Recent Observations* Vol. 193 (eds Eracleous, M. et al.) 103 (ASP Conf. Ser., 1996).
13. Bickel, G. V. & Begelman, M. C. *Astrophys. J.* **467**, 597–621 (1996).
14. Mirabel, I. F. & Rodrigues, L. F. *Nature* **392**, 673–676 (1998).

Cell adhesion

New way to activate caspases

Erkki Ruoslahti and John Reed

Most cells need to be attached to a self-produced external meshwork known as the extracellular matrix in order to survive — without such anchorage they undergo apoptosis¹. The cells attach through integrins, heterodimeric membrane proteins that link the intracellular cytoskeleton with the extracellular matrix. Many of the 25 or so known integrins recognize a tripeptide, arginine–glycine–aspartate (or RGD in single-letter code), in target proteins of the extracellular matrix². And on

page 534 of this issue, Buckley *et al.*³ provide a startling new interpretation for the role of these RGD peptides in apoptosis.

In vitro experiments have shown that small peptides containing the RGD motif bind to integrins and, when presented to a cell in a soluble form, inhibit cell attachment. According to current thinking, the peptide-blocked integrins cannot provide the signals that cells would receive from matrix-bound integrins, resulting in changes in cell shape and, ultimately, apoptosis^{1,4}. Because RGD-

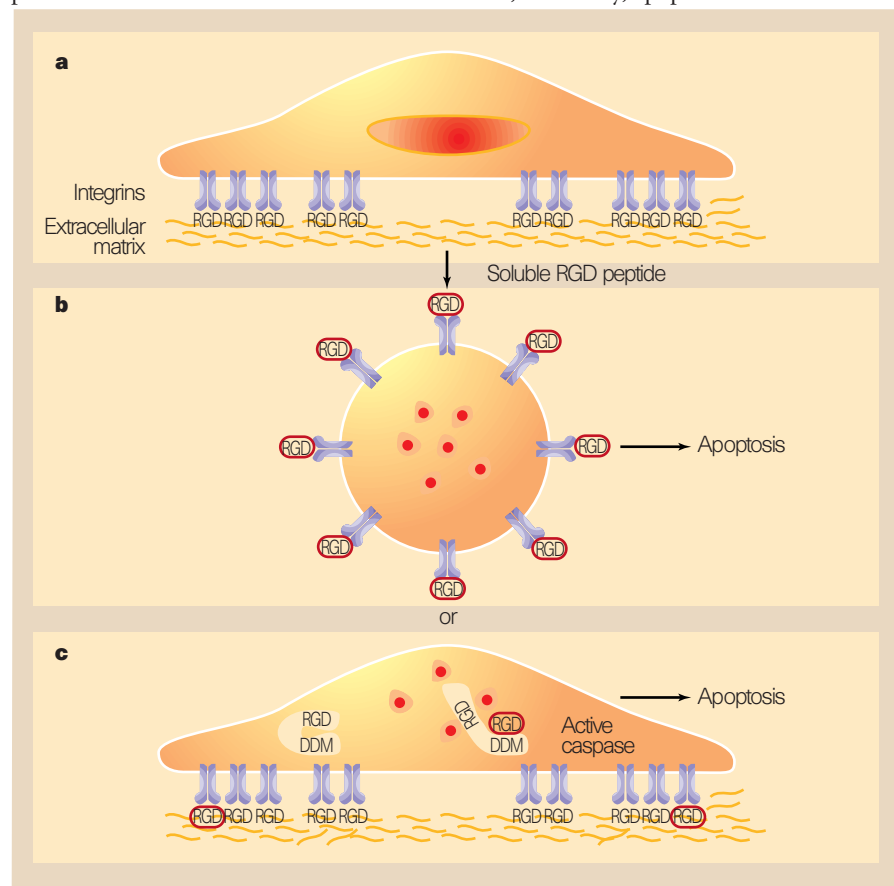


Figure 1 The arginine–glycine–aspartate (RGD) peptide in apoptosis. a, Many cells attach, through integrins, to a substrate such as the extracellular matrix in order to survive. b, Addition of soluble RGD peptides causes apoptosis. One model is that the soluble peptides block integrin signalling to the extracellular matrix, so the cell detaches and the missing integrin signal causes apoptosis. c, Buckley *et al.*³ propose a new explanation. They show that soluble RGD peptides cause apoptosis by activating procaspase-3 inside the cell. The RGD peptide may out-compete an RGD motif in the procaspase, blocking an intramolecular interaction with the sequence DDM and causing the procaspase to open up and be activated.

based anti-thrombotic drugs are already in clinical use, and RGD drugs to treat osteoporosis and inhibit angiogenesis (the development of new blood vessels) are under development², knowing how these peptides act is of considerable importance.

The mechanism proposed by Buckley *et al.*³ is that the soluble RGD peptide activates a caspase in the cytoplasm of the cell. Caspases are proteases that are critical in apoptosis. They amplify and propagate apoptotic signals and execute the apoptotic programme by degrading intracellular proteins vital for cell survival⁵. So, activation of caspases would be an alternative explanation for the pro-apoptotic activity of RGD peptides. The authors note that there is an RGD sequence near the active site of caspase-3 — the caspase that is reportedly activated by RGD peptides. They surmise that this RGD site is normally engaged in an intramolecular interaction, which keeps the procaspase inactive. But a soluble RGD peptide could block this intramolecular interaction, leaving the procaspase RGD site free. The procaspase could then open up (see Fig. 1) and be activated by, for example, self-digestion.

Supporting this proposed mechanism, the sequence aspartate-aspartate-methionine (DDM) — which resembles an RGD binding site in integrin β -subunits⁶ — is found in a loop region of procaspase-3 that is proteolytically cleaved during activation. Unfortunately, the authors did not test whether the DDM peptide can activate caspases, nor did they mutate the DDM site in caspase-3 to test their hypothesis (mouse caspase-3, for example, contains the sequence glutamate-glutamate-methionine, EEM). Although the crystal structure of active caspase-3 is known⁷, the structure of the unprocessed pro-form is needed to confirm the speculation about an intramolecular interaction between RGD and DDM.

Buckley *et al.*³ have shown that RGD peptides specifically activate procaspase-3 isolated from cells by immunoprecipitation, and experiments using recombinant, purified procaspase-3 should be useful in working out the mechanism. One puzzling result from the experiments with intact cells is that integrins do not seem to be needed for internalization of RGD peptides — a RAD peptide, which does not bind to integrins², was also internalized by the cells. Integrins tend to transport ligands into the cytoplasm; viruses and intracellular bacteria use integrins as receptors, and gene therapists build viruses that do the same². On this basis, we would expect only the RGD peptide to find its way into the cytoplasm.

Intriguing as this new mechanistic insight into the RGD peptide is, it is unlikely to be the whole story. For example, it is difficult to explain the pro-apoptotic activity of anti-integrin antibodies³ on this basis. This activity seems to be tied to the specificity of the anti-

bodies for an individual integrin, and not all of the antibodies contain an RGD sequence. It seems more plausible that these antibodies act through extracellular inhibition of cell attachment, rather than through direct activation of caspases within the cell. But there may be a way to reconcile the two mechanisms. Perhaps, when cells become detached from the extracellular matrix, RGD-containing proteins that would normally be used to replenish the matrix are degraded, and generate enough RGD peptides for caspase activation. In this regard, it would be informative to test the effect of anti-integrin antibodies on lymphocytes, which do not require attachment and do not make a matrix.

Buckley and colleagues' results open other fascinating possibilities. Could the HIV Tat protein, which is taken up by cells and contains an RGD sequence in one of its permutations, activate caspases? Could the anti-angiogenic compounds angiostatin and endostatin, for which no cell-surface receptor has been found, also act in this manner? Could RGD peptides with different integrin-binding specificities² differ in their abilities

to activate individual caspases? (Buckley *et al.* used RGD peptides that bind to all RGD-directed integrins.) Could degradation of proteins from the extracellular matrix generate enough RGD peptide to cause apoptosis of cells in an involuting tissue, such as breast tissue that has stopped lactating⁹? Finally, can the RGD-mediated activation of caspases be exploited to trigger apoptosis selectively in cancer cells or inflammatory cells, without toxicity? The new work will stimulate much research along these lines. □

Erkki Ruoslahti and John Reed are at The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA.

e-mails: ruoslahti@burnham-inst.org
jreed@burnham-inst.org

1. Frisch, S. M. & Francis, H. J. *Cell Biol.* **124**, 619–626 (1994).
2. Ruoslahti, E. *Annu. Rev. Cell Dev. Biol.* **12**, 697–715 (1996).
3. Buckley, C. D. *et al. Nature* **397**, 534–539 (1999).
4. Chen, C. S., Mrksich, M., Huang, S., Whitesides, G. M. & Ingber, D. E. *Science* **276**, 1425–1428 (1997).
5. Salvesen, G. S. & Dixit, V. M. *Cell* **91**, 443–446 (1997).
6. Pasqualini, R., Koivunen, E. & Ruoslahti, E. *J. Cell Biol.* **130**, 1189–1196 (1995).
7. Rotanda, J. *et al. Nature Struct. Biol.* **3**, 619–625 (1996).
8. Brooks, P. C. *et al. J. Clin. Invest.* **96**, 1815–1822 (1995).
9. Werb, Z. *et al. Kidney Int.* **54**, 68–74 (1996).

Photochemistry

Electronic motion in DNA

Mark Ratner

DNA's biological task involves information storage and propagation. To perform its function effectively, DNA must be stable, robust and repairable. Unlike proteins such as cytochromes and the photosynthetic reaction centre, DNA is not primarily an electron-transfer species. Still, the structure of DNA with its π -electron system of four bases stacked upon each other is reminiscent of certain molecular metals, such as

doped phthalocyanines, so the possibility of long-range charge transfer is an intriguing one. Consequently, there has been a series of tantalizing reports on electron transfer in DNA. A new paper in the *Journal of the American Chemical Society* by Meggers *et al.*¹ reveals some regularities in the step-wise motion of electron 'holes' in the DNA structure, measures the length dependence of such transfers, and makes some general

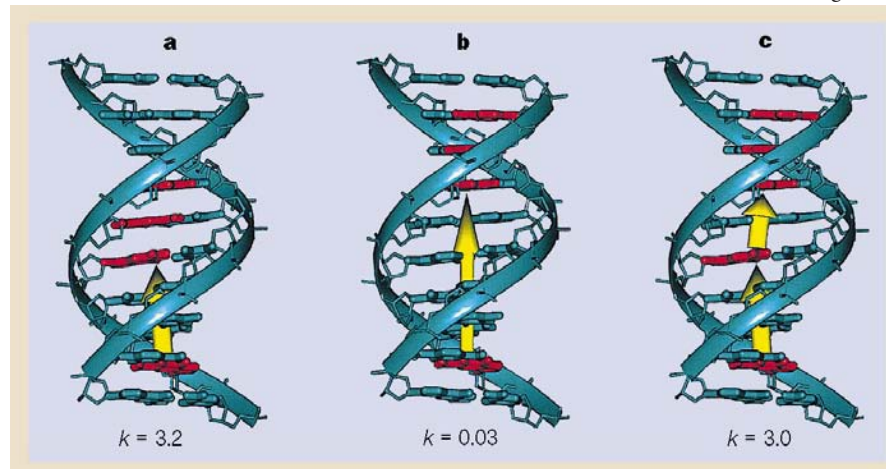


Figure 1 Conduction in double-stranded DNA measured by Meggers *et al.*¹. In a, an electron 'hole' passes from a guanine in a guanine–cytosine (GC) base pair (red) to another G by way of superexchange, jumping over two intermediate adenine–thymine (AT) pairs (blue). The rate of hole transfer k is shown below. In b, the superexchange transports the hole twice as far, but the transfer is 100 times as slow; whereas in c, two short superexchange hops by way of an intermediate G cover the same distance as in b, but as quickly as in a. The transfer from one G to the next occurs coherently (superexchange), but the transport over several G bases occurs incoherently (hopping).

arguments concerning the mechanism of charge transfer in the DNA stack.

The fundamental mechanism of molecular electron transfer requires an electronic mixing pathway between the initial and final states. In most intramolecular and intermolecular electron-transfer reactions, the charge is localized only at the first and last sites: the charge is thought to move between these two points in a single, coherent jump — much like a kicked football². But when the states between the initial and final ones are sufficiently low in energy the charge can stop along the way, so that its overall trajectory looks more like the path of a wandering drunk. The coherent transfer process cannot get very far at room temperature, because the orbitals in which the electron density is found do not extend effectively over long distances and thermal disorder further localizes the charge. The latter, diffusive motion is responsible for the ordinary conductivity of real metals^{3–5} (as described by Ohm's law).

So, depending on the energy, one expects different mechanistic behaviour: if the 'bridging states' (in DNA, these are the intervening base pairs) are very high in energy compared with the initial and final states, we should see coherent transport. This is generally called superexchange, and is characterized by a rapid exponential decay of the transfer rate or yield as a function of distance². But if the intermediate bridging states are comparable in energy, or lower in energy, than the initial state, then one expects to see incoherent, hopping behaviour that decays only slowly with distance.

Meggers *et al.*¹ have used a clever photochemical method for inducing an electron hole on a guanine (G) base in the DNA structure. This hole (a positive charge centre prepared by removing an electron from the DNA stack) then wanders along the DNA chain. The energy of the hole when residing on adenine, cytosine or thymine bases is substantially higher than when on G, so the electron never stops except on other G bases. By measuring DNA fragments produced by chemical cleavage at the different G sites, Meggers *et al.* can actually measure the probabilities, and therefore the relative rates, of hole transfer along the strand.

The results are striking. They suggest that the electron does indeed hop incoherently among the G bases in a kind of random walk. This gives a weak, algebraic decay of the transfer rate with length. When moving from one G base to the next, however, the electron cannot stop in mid-journey, so that the transfer between two different G bases is like coherent superexchange, and decays exponentially with distance. Figure 1 shows three different hole transfer situations, and their measured rates.

The suggestion of two different mechanisms for superfer, one corresponding to coherent superexchange and the other to a

random walk, reflects recent discussions^{3–6} and observations for the photosynthetic reaction centre⁷ (where the intermediate site is the bridging bacteriochlorophyll) and in synthetic donor/bridge/acceptor intramolecular electron-transfer systems⁸. It implies that, although superexchange could be appropriate for short-range electron transfer in DNA^{9–12}, for effective long-range transfer only the hopping process can occur. This is similar to what happens in both metals and molecular conductors; it differs from the (high-field) coherent transport described in carbon nanotubes¹³.

The robust, malleable, one-dimensional structure of DNA is unique. It can be used to design functional nanostructures, and its charge transport capability in the appropriate energy regime can be quite good. The report by Meggers *et al.* and other recent measurements and models for charge motion in DNA may help settle a controversial question about the conductivity of this biologically important molecule¹². They suggest that DNA's unique structure and π -electron bases do indeed provide appropriate pathways for long-range charge transport, but that the mechanisms for long-range transport and short-range transfer differ entirely. Once charges (especially holes) are created on the DNA chain, then hopping-type conduction can apparently occur among the G sites; this is observed by Meggers *et al.*, and was also suggested by earlier workers^{14,15}. This simplistic hopping mechanism would indeed make DNA a fairly good hole conductor. The actual conductivity, however, will vary with the location and density of G-sites, with the injection energy and probability for the holes, and with the level of disordering in the helical stack. These sensitivities help to explain the wide variety of conductivities suggested by previous experiments and models. □

Mark Ratner is in the Chemistry Department, Northwestern University, Evanston, Illinois 60208, USA.

e-mail: ratner@chem.nwu.edu

1. Meggers, E., Michel-Beyerle, M. E. & Giese, B. *J. Am. Chem. Soc.* **120**, 12950–12955 (1998).
2. Marcus, R. A. & Sutin, N. *Biochim. Biophys. Acta* **811**, 265–322 (1985).
3. Davis, W. B. *et al.* *J. Phys. Chem.* **101A**, 6158–6164 (1997).
4. Skourtis, S. & Mukamel, S. *Chem. Phys.* **197**, 367–388 (1995).
5. Pollard, W. T., Felts, A. K. & Friesner, R. A. *Adv. Chem. Phys.* **93**, 77–134 (1996).
6. Jortner, J. *et al.* *Proc. Natl Acad. Sci. USA* **95**, 12759–12765 (1998).
7. Bixon, M. & Jortner, J. *J. Chem. Phys.* **107**, 5154–5170 (1997).
8. Davis, W. B. *et al.* *Nature* **396**, 60–63 (1998).
9. Turro, N. J. & Barton, J. K. *J. Biol. Inorg. Chem.* **3**, 201–209 (1998).
10. Lewis, F. D. *et al.* *Science* **277**, 673–676 (1997).
11. Krider, E. S. & Meade, T. J. *J. Biol. Inorg. Chem.* **3**, 222–225 (1998).
12. Beratan, D. N., Priyadarsky, S. & Risser, S. M. *Chem. Biol.* **4**, 3–8 (1997).
13. Tans, S. J. *et al.* *Nature* **394**, 761–764 (1998).
14. Razskazovskii, Y. *et al.* *J. Phys. Chem. B* **101**, 1460–1467 (1997).
15. Gasper, S. M. & Schuster, G. B. *J. Am. Chem. Soc.* **119**, 12762–12769 (1997).

Daedalus

Cell squeezing

We get many useful substances from plants. Sugar, digitalis, nicotine, turpentine, caffeine, essential oils, alginates, steroids — the list is endless. Usually, the plant must be grown, harvested and processed to extract the product. In theory, the plant tissues could be grown as a cell culture; even then, extraction would be a nuisance. Only a few single-cell products, such as penicillin and alcohol, are freely released by the growing cells. Daedalus now has a new twist.

Supercritical fluids, he points out, are wonderful solvents, with very high molecular diffusivity. Some of them (such as carbon dioxide, nitrous oxide and xenon) have critical points near ambient. Single cells are incompressible and can withstand great hydrostatic pressures. So Daedalus is developing supercritical cell culture.

Carbon dioxide, the essential carbon source for all green plants, seems the ideal supercritical medium. Yet such a high concentration could be damaging; xenon with a dash of carbon dioxide might be safer. But whatever mixture turns out best, cell culturing will be transformed. Its biochemistry will be speeded up enormously: feedstock and product molecules will diffuse in and out through the cell walls at a great rate. Even a small culture will churn out pharmaceuticals, alkaloids or perfumes in copious quantity. Existing culturing methods, such as those for antibiotics, and proteins from modified *E. coli*, will also go supercritical.

An older trade, herbal medicine, should also benefit. Traditional prescriptions can be highly complex for a very vague claimed action — typically 'clearing toxins' (unspecified) or 'boosting the immune system'. Their active ingredients, if any, and how they achieve their alleged effects, are seldom known. But supercritical culturing could generate plant metabolites in such quantities that their benefits could be swiftly clarified.

Even genetic engineering might be speeded up. In supercritical conditions, plasmids and proteins should drift in and out of cells, and be exchanged between them, with unprecedented ease. Indeed, a supercritical mixed-cell culture might even act as a sort of hybrid super-organism. All its different cell types, whether from a single organism or even from disparate species, might pool their biochemical and genetic resources into a mighty synergistic living combine, a cellular Frankenstein's monster.

David Jones

Obituary

Jean Leray (1906–98)

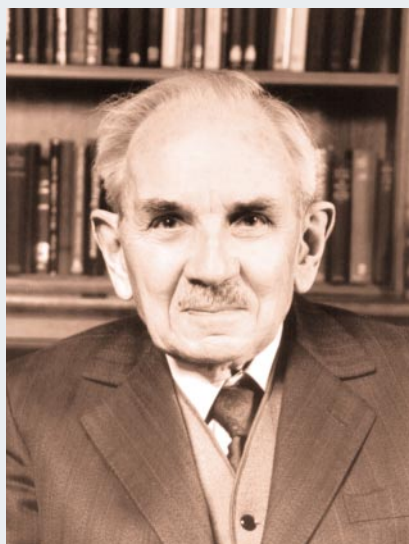
Master of applied mathematics

Only a few months after the death of André Weil, whose work and influence was described here on 29 October last year (*Nature* 395, 848; 1998), another major figure has left the mathematical scene. Jean Leray died on 10 November, at La Baule in France, after a long career spent mostly as a professor at the Collège de France.

Weil and Leray had curiously parallel lives: both were born in 1906, both were educated at the Ecole Normale Supérieure in Paris, and both died in 1998. Weil left France during the Second World War and spent the remainder of his career in Princeton, while Leray was captured with his combat unit in 1940 and was detained in Austria until the end of the war, when he returned to France. They symbolize the two opposing trends that have characterized mathematics over the past 50 years.

Leray viewed mathematics as a tool for modelling, and drew his inspiration from problems in mechanics and physics, such as fluid dynamics and wave propagation. He was fond of explaining how the road from mathematics to applications is two-way, and how a purely mathematical theorem (concerning, for instance, the existence and uniqueness of solutions of systems of partial differential equations) might have profound physical implications.

Leray's studies of the Navier–Stokes equations, which are fundamental to fluid dynamics, were of this kind. In an epoch-making paper of 1934, published in *Acta Mathematica*, he showed that smooth initial data may give rise to non-stationary solutions which remain smooth for a finite time only, after which they become irregular (the maximum velocity 'blows up'), and satisfy the Navier–Stokes equations only in a generalized sense; that is, such solutions do not satisfy the equations at every point, but must be integrated against smooth test functions. Leray termed such solutions 'turbulent'. In the same paper, he proved that regular solutions are uniquely determined by their initial data, and suggested that, on the other hand, there might be many 'turbulent' solutions corresponding to the same initial data. This conjecture remains unproved, and is one of the competing arguments for describing the onset of turbulence.



Weil, on the other hand, viewed mathematics as a structure developing according to its own needs, and prided himself on never having done anything applied. He was one of the founders of the Bourbaki group, which set itself the task of rewriting mathematics on firm premises, free from all compromises with common sense and so-called geometrical or physical intuition. This line of thinking was to be enormously influential in France, where it came to dominate teaching and research until the mid-1970s. Leray kept away from the crowd and pursued his own brand of mathematics, firm in his conviction that it was the right one, and that it would prevail. I still remember his lectures. He was a mild-mannered, dapper man with a grey moustache, who squinted at his audience and lost it rather quickly; but he continued to write on the blackboard amidst a respectful silence, confident that the mathematics were there for all to see and needed no further explanation.

By great irony, Leray himself turned out to be one of the main sources for the Bourbaki brand of mathematics. During the five years he spent as a prisoner of war in Oflag XVIII, he not only helped found a 'university' but also pursued his own research. He was however an ardent patriot, and he did not want the Germans to discover his competence in fluid mechanics, lest he be compelled to contribute to their war effort. So he turned to 'useless' mathematics, namely algebraic topology, the study of algebraic structures associated with topological spaces.

The topic did not catch the interest of German intelligence. But it has become central to modern geometry and physics, because in both cases one is concerned

with symmetries, that is, with the way groups can act on an underlying space: the description of these actions, and the way the orbits fit in to reconstruct the whole space, are problems that cannot be tackled without the tools of algebraic topology. Two of these tools are products of Leray's war years: the 'spectral sequence', which has been central in the investigation of homotopy groups of spheres (the different ways to map a sphere into a sphere of lower dimension); and the 'theory of sheaves', which is fundamental to many areas of pure mathematics, such as the study of functions of several complex variables.

After the war, Leray returned to the study of partial differential equations, and remained highly active until his death. Three volumes of his selected papers, edited by Paul Malliavin, were published in 1997 by Springer-Verlag and the Société Mathématique de France, and the introductions to each volume give more detail about his work. Here, I would like to stress Jean Leray's modernity. His 1934 paper on the Navier–Stokes equations was truly 40 years ahead of its time, and is unsurpassed even today. He solved nonlinear systems of partial differential equations at a time when the tools for solving linear ones had yet to be invented. Not until the 1960s did the mathematical world master the notion of weak solutions, Leray's 'turbulent' solutions, and not until the 1970s were his methods for solving nonlinear systems fully understood.

Leray was so far ahead of his time because of his tremendous technical capability and geometrical insight. In his hands, energy estimates for partial differential equations became combined with ideas from algebraic topology (such as fixed-points theorems) in a highly original combination which cracked open the toughest of problems. He was the first to adopt the modern point of view, whereby a function is not a complicated relation between two sets of variables, but a point in some infinite-dimensional space; and he was the first to solve problems in functional calculus (and hence, partial differential equations) by using the geometry of such spaces. All of these ideas are classics, and little has since been added to them. If we define algebra as the study of abstract mathematical structures, and analysis as solving equations arising from physics or other sciences, then Weil can be said to have been the first modern algebraist and Leray the first modern analyst.

Ivar Ekeland

Ivar Ekeland is at the University of Paris-Dauphine, 75775 Paris Cedex 16, France. e-mail: ivar.ekeland@dauphine.fr

Making make-up in Ancient Egypt

The extensive use of green, white and black make-up has been known since the earliest periods of Egyptian history^{1,2}. We have investigated cosmetic powders dating from between 2000 and 1200 BC that were preserved in their original containers. Quantitative crystallographic and chemical analysis of the organic and mineral components of the powders enabled us to identify two natural lead-based compounds: crushed ore of galena (PbS) and cerussite (PbCO₃). We also found two unexpected constituents: laurionite (PbOHCl) and phosgenite (Pb₂Cl₂CO₃). Because they are neither natural extracted ores nor products resulting from subsequent ageing or chemical modification, laurionite and phosgenite appear to be synthetic products manufactured by the Egyptians using 'wet' chemistry.

The cosmetic powders were exceptionally well conserved in containers made of stone (alabaster, haematite, marble), ceramic, wood and reed, which are housed in the Egyptian department of the Louvre Museum in Paris³. The powders contained the lead chloride compounds laurionite and phosgenite (Fig. 1a), which are both very rare in nature. They are found as corrosion products of lead artefacts⁴ or in lead slags from ancient silver-mining operations that have been weathered by sea water⁵.

Phosgenite is naturally formed by the oxidation of minerals containing lead, but only in those few locations where the ore comes into contact with carbonated and chlorinated waters. Provided that such natural products were extracted, their abundance is far too small to account for their extensive use in cosmetics over a period spanning at least eight centuries (Table 1). Even if running water provided chlorides, chemical weathering of the powders in their original containers is also an unlikely source of these compounds, as the reed containers are very well preserved and some have intact hieroglyphics in black ink describing their contents. Moreover, no foreign cations or identified chlorinated phases could be detected in the powders, and the coarse galena grains have cleaved faces that are free of any surface damage

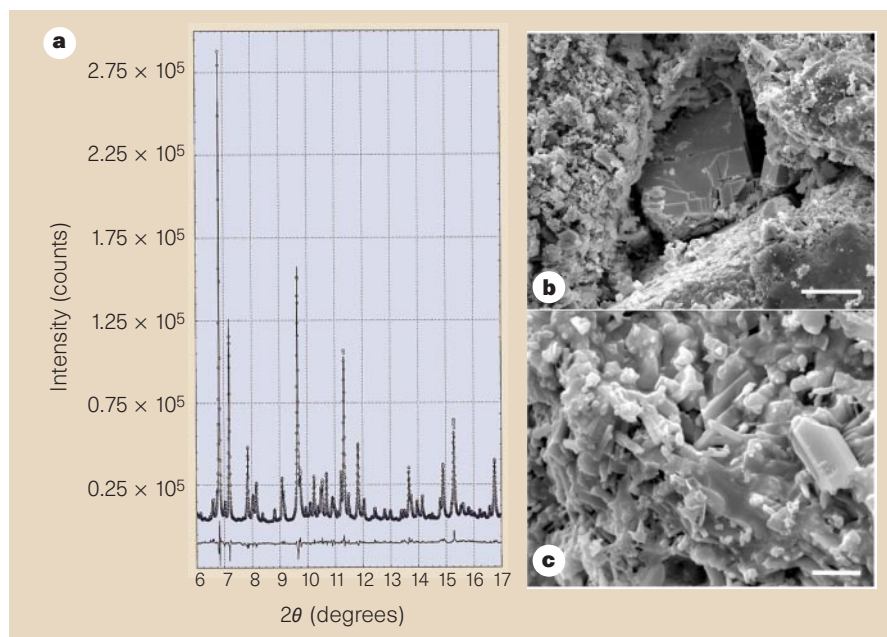


Figure 1 Analysis of cosmetic powders. **a**, X-ray diffraction patterns for powder sample 4. Circles, experimental data points; top trace, calculated by Rietveld refinement; bottom trace, difference (observed minus calculated). **b**, Scanning electron micrograph (SEI) of sample 5. Compounds of lead and chloride are aggregated with cubic crystals of galena that have well developed faces free of alteration. Scale bar, 20 μm. **c**, SEM of sample 5. Laurionite grains have a fine morphology, Scale bar, 2 μm.

(Fig. 1b). The alteration of natural lead minerals within the make-up is therefore unlikely. The presence of large amounts of laurionite and phosgenite in the Egyptian cosmetics cannot therefore be explained by weathering or alteration. This evidence indicates that laurionite and phosgenite were synthesized artificially.

Support for this statement comes from recipes of medicinal products reported by classical authors^{6,7} in which purified silver foam (PbO) was crushed and mixed in water with rock salt and sometimes natron (mainly Na₂CO₃) and then filtered; the procedure was repeated every day for several weeks. We repeated these chemical reactions in our laboratory by mixing PbO and NaCl powders in carbonate-free water. The corresponding slow reaction produces an alkaline solution that can be maintained at a neutral pH to simulate the daily replacement of water. The end precipitate was then

identified as laurionite by X-ray diffraction. Analysis of scanning electron micrographs reveals that our crystals and the Egyptian ones (Fig. 1c) have a similar fine morphology. It should be pointed out that, in the presence of carbonates, the reaction produces phosgenite⁴.

Taken together, these results indicate that laurionite and phosgenite must have been synthesized in Ancient Egypt using wet chemistry. The Egyptians manufactured artificial lead-based compounds, and added them separately to the cosmetic product. The underlying chemical reactions are simple, but the whole process, including many repetitive operations, must have been quite difficult to achieve.

Analysis of the organic constituent of the powders shows the presence of significant amounts of fatty acids. A variable amount (up to 7%) of lipids, identified by gas chromatography, was presumably added to the dry powder to yield a variety of textures (see Supplementary Information). The choice of the original minerals, the subsequent chemical process, the variable proportions of the mineral components, and the addition of lipids enabled the Egyptians to produce a wide range of cosmetics for specific purposes. Hieroglyphic inscriptions found on reed receptacles kept in a wooden box (from the XVIIIth dynasty) make a distinction between the one containing the black make-up 'mesdemet' and

Table 1 Proportion of the four main mineral phases

Sample	Galena	Cerussite	Phosgenite	Laurionite
1	100			
2	50	13	37	
3	28	48	24	
4	43	27	29	1
5	12		72	16
6	62	28		10
7	24	25	16	35

Data were obtained by Rietveld refinement of the powder diffraction patterns. The samples were taken from containers made of alabaster (samples 1, 5, 6 and 7) and reed (samples 2, 3 and 4). See Supplementary information for a more detailed description.

two others with “eye lotion to be dispersed, good for eyesight”. Analysis⁸ has shown that the former product was galena and the latter were mixtures of lead chloride and carbonate. Moreover, Egyptian medical papyrus lists one hundred recipes for treating problems of the eyelids, iris and cornea, as well as trachoma and conjunctivitis. Some of these materials have been identified as malachite, galena and their derivatives, although several others have not yet been identified⁹.

Fire-based technology has been used to manufacture Egyptian blue pigments since about 2500 BC (refs 1,2). We have now shown that wet chemistry was used as long ago as 2000 BC. This finding shows that the chemical technology of Ancient Egypt was more sophisticated than previously thought.

P. Walter*, **P. Martinetto***, **G. Tsoucaris***,
R. Bréniaux†, **M. A. Lefebvre†**, **G. Richard†**,
J. Talabot†, **E. Dooryhee‡**

*Laboratoire de Recherche des Musées de France,
UMR 171-CNRS, 6 rue des Pyramides,
75041 Paris Cedex 1, France
e-mail: p.walter@culture.fr

†L'Oréal-Recherche,
90 rue du Général Roquet,
92583 Clichy Cedex, France

‡European Synchrotron Radiation Facility,
BP 220, 38043 Grenoble Cedex, France

- Forbes, R. J. *Studies in Ancient Technology* Vol. III (Brill, Leiden, 1955).
- Lucas, A. & Harris, J. R. *Ancient Egyptian Materials and Industries* 80–97 (Arnold, London, 1963).
- Vandier d'Abbadie, J. *Catalogue des objets de toilette égyptiens du Musée du Louvre* (Réunion des Musées Nationaux, Paris, 1972).
- Edwards, R., Gillard, R. D., Williams, P. A. & Pollard, A. M. *Mineral. Mag.* **56**, 53–65 (1992).
- Conophagos, C. in *The Evolution of the First Fire-Using Industries* (eds Wertime, T. A. & Wertime, S. F.) 181–191 (Smithsonian Institution Press, Washington DC, 1982).
- Wellman, M. *Dioscoridis Pedanii, de Materia Medica, Libri Quinque* (Weidmannsche, Vienna, 1958).
- Zehnacker, H. *Plin l'Ancien, Histoire Naturelle XXXIII* (Belles Lettres, Paris, 1983).
- Lucas, A. in *Quibell Ann. Serv. Antiq. Egypte* **2**, 141–143 (1901).
- Bardinet, T. *Les papyrus médicaux de l'Égypte pharaonique* (ed. Fayard, Paris, 1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Josephson effect and a π -state in superfluid ^3He

Evidence has been reported for Josephson-type behaviour in the flow of superfluid helium-3 (^3He) through an array of 4,225 apertures^{1,2} and for a metastable π -state at low temperature³. Here we show that an array of holes cannot be expected to act as a single weak link in such a system, and that these results can be explained without invoking the superfluid analogue of the much-debated ‘ π -junction’ in other systems.

Convincing experimental evidence for the Josephson effect in superfluid ^3He was first reported ten years ago^{4,5}. The mass current J and the quantum-mechanical phase difference $\Delta\phi$ across a single microscopic orifice are well represented by the relations $J = J_c \sin \zeta$, and $\Delta\phi = \zeta + \alpha \sin \zeta$, where J_c is the critical current, ζ is an auxiliary phase angle and α is a non-ideality parameter. These relations describe both the ideal Josephson regime, $\alpha \approx 0$, observed close to the superfluid transition temperature, and the phase-slippage regime, $\alpha > 1$, which occurs at lower temperatures. This relation is represented by the solid line in Fig. 1a for $\alpha = 2$. In this hysteretic regime, the current is double-valued around $\Delta\phi = \pi$. In an array of holes, if half of the holes sit on the upper branch, J_+ , and the other half sit on the lower one, J_- , the measured current through the array is $(J_+ + J_-)/2$, as represented by the dotted line in Fig. 1a. Numerical simulations indicate that a suitable excitation drive can set an array of holes into this state.

The time evolution of the membrane displacement, X , monitoring the flow through an array of N apertures, is obtained by solving numerically the set of equations

$$\Delta\phi = \zeta_i + \alpha_i \sin \zeta_i, \quad (1)$$

$$dX/dt = C \sum_{i=1}^N \sin \zeta_i, \quad (2)$$

$$\frac{d(\Delta\phi)}{dt} = V \sin(\omega t) - X - \frac{1}{\omega_0 Q} \frac{dX}{dt} \quad (3)$$

where C is related to the small signal oscillation frequency, $\omega_0^2 = C \sum_{i=1}^N (1/\alpha_i)$, $V \sin(\omega t)$ is the external drive of maximum amplitude V at frequency ω , and Q is the quality factor of the resonator.

The result of such a simulation is shown in Fig. 1b. We used 200 holes with a gaussian distribution of α_i values with mean value 2 and standard deviation 0.1. Apertures in different regions of the array experience different series hydraulic inductions, and the spread in α_i in the previous experiments^{1–3} could actually be larger. The Q value of the resonator was set at 1,000. The external drive, $V \sin(\omega t)$ in equation (3), was made of four periods at a frequency $\omega = 1.1\omega_0$ and the amplitude was adjusted until a stable low-amplitude state was reached. In this state, about half of the holes have experienced a phase slip. The kinetic energy is then stored in the circulating currents in the array and a small perturbation can drive the system out of this state, in one direction or in the other, as shown in Fig. 1b for a positive perturbation.

A plot of $d(\Delta\phi^*)/dt$, defined as $-X(t)$, versus $\Delta\phi^*$, defined as $-\int_0^t X(t)dt$, is represented in Fig. 1c. According to equation (3), these $\Delta\phi^*$ are not the true $\Delta\phi$, but such a procedure was used to plot Fig. 3a in ref. 3,

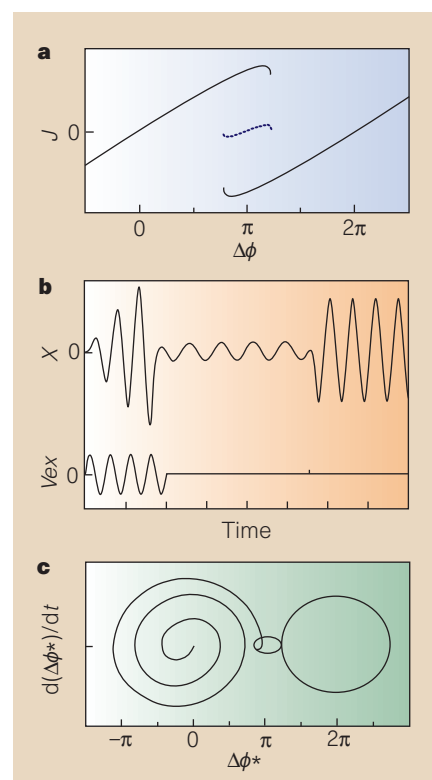


Figure 1 Numerical simulations. **a**, Current-phase relationship for $\alpha=2$. Dotted line, artefact of the measurements in an array of holes discussed in the text. **b**, Time evolution of the membrane amplitude X in response to an external drive V_{ex} . **c**, Corresponding trajectory in the $d(\Delta\phi^*)/dt$ versus $\Delta\phi^*$ plane.

with which our numerical data must be compared. The agreement with real data is striking, taking into account that there is neither noise nor filtering in the simulated data.

With these input parameters, the excitation window leading to a stable low-amplitude state is very narrow. Slight variations in the distribution of α_i , in the shape of the applied drive, or in the details of the current-phase relationship around π could widen this window.

In conclusion, our analysis shows that direct measurements of the current-phase relationship in an array of holes should be interpreted with care. The data in refs 1–3 can be reproduced by simple numerical simulations that involve only physical mechanisms previously demonstrated to take place in a single aperture. By themselves, these data do not reveal the existence of an exotic π -junction, hidden degrees of freedom, or any influence of textural effects.

Olivier Avenel*, **Yury Mukharsky***,
Eric Varoquaux†

*CEA-DRECAM,
Service de Physique de l'État Condensé,
Centre d'Etudes de Saclay,
91191 Gif-sur-Yvette Cedex, France
e-mail: avenel@spec.saclay.cea.fr
†CNRS-Laboratoire de Physique des Solides,

Bât. 510, Université Paris-Sud,
91405 Orsay, France

Backhaus et al. reply — Avenel *et al.* have suggested a mechanism that might explain the recently discovered¹ metastable π -state in a superfluid ³He weak-link array. We are pleased that our experiment is leading to new ideas that may extend the understanding of weak-link arrays. We agree with Avenel *et al.*'s comment that, when the individual apertures are in a (short coherence length, low temperature) hysteretic regime, collective phenomena quite distinct from single weak-link behaviour might be observed. Nevertheless, in the temperature regime in which the coherence length is comparable to the aperture dimensions, we have shown that the collective behaviour of the array is similar to that of a single weak link^{2,3}.

Although we are gratified that the mechanism proposed by Avenel *et al.* for the π -state expands on our previous conjecture¹ that internally trapped circulating currents could be involved in its existence, we do not believe that it is the only possible explanation. This is because agreement between a given data set and a numerical simulation, using a model with several adjustable parameters, represents a check that the model is consistent with the data, but does not qualify as a proof of the model⁴. Therefore, we cannot say to what extent their model represents physical reality better than other theories of the π -state that have been recently proposed⁵.

However, several aspects of their simulation disagree with our experimental observations. In particular, the model presented by Avenel *et al.* does not seem to conserve energy in the oscillator, contrary to our experimental results. Their model predicts that the $I(\phi)$ relation should extend beyond $\phi=\pi$ before the onset of the π -state. In contrast, the collapse into the π -state occurs at $\phi < \pi$. Finally, as shown in their Fig. 1c and supported by our own simulations, their model leads to metastable states at positions other than π , a feature also in contradiction to the data.

We hope that Avenel *et al.* will continue to refine their simulations to make predictions that could lead to a conclusive test of their underlying model.

**S. Backhaus, R. W. Simmonds, A. Loshak,
J. C. Davis, R. E. Packard**

Department of Physics,
University of California, Berkeley,
Berkeley, California 94720, USA

1. Backhaus, S. *et al.* *Nature* **392**, 687–690 (1998).
2. Pereversev, S. V., Loshak, A., Backhaus, S., Davis, J. C. & Packard, R. E. *Nature* **338**, 449–451 (1997).
3. Backhaus, S., Pereversev, S., Loshak, A., Davis, J. C. & Packard, R. E. *Science* **278**, 1435 (1997).
4. Varoquaux, E., Avenel, O., Ihas, G. & Salmelin, R. *Physica B* **178**, 309–317 (1992).
5. Hatakanaka N. *Phys. Rev. Lett.* **81**, 17 (1998).

How and why a parasitic nematode jumps

Jumping is an unusual behaviour performed by some nematode species¹, but has been seen only in the infective or phoretic stages of species associated with insects^{1–3}. This correlation suggests that jumping may be involved in the location of insect hosts. We find that infective juveniles of the insect-parasitic nematode *Steinernema carpocapsae*, when standing on their tails, are triggered to jump by the presence of host-associated volatile cues, and that they tend to jump towards them. Directional jumping in response to information about insect proximity could be an adaptation for host attack by this parasite.

S. carpocapsae is a lethal endoparasite capable of infecting a broad range of insect species⁴. The infective juvenile, which emerges from a depleted host cadaver to seek out a new insect to infect, is the only free-living stage. Infective juveniles tend to be found at the soil surface⁵, and use an ambushing search strategy in which they stand on their tails and attach to a passing insect². They are more likely to be picked up by an insect when they are standing because of the reduced surface tension holding them to their substrate.

Jumping, which previously was inaccurately described¹, is initiated when a standing nematode quickly bends the anterior half of its body until its head region makes contact with the ventral side of its body. The two body regions appear to be held together by the film of water covering the nematode's body¹. The nematode now has resistance to the bending of its body, so it can use its normal sinusoidal crawling behaviour to slide its body in a posterior direction, causing the loop to become progressively smaller and the bend in its body to become more acute.

Eventually, the body becomes so contorted that the cuticle on the dorsal side becomes extremely stretched and the cuticle on the ventral side kinks, generating sufficient force to break the surface tension forces holding the two body parts together. As its body straightens out, enough force is applied to break the surface tension forces holding the nematode to the substrate and propel it through the air. The forces generated by this jumping mechanism are sufficient to propel nematodes an average distance of 4.8 ± 0.8 mm (nine times the nematode's body length) and an average height of 3.9 ± 0.1 mm (seven times the nematode's body length).

If jumping behaviour has been acted on by natural selection to function in finding hosts by ambush, then we might expect it to

be triggered by the presence of potential hosts, and for nematodes to jump towards the host. We found that *S. carpocapsae* infective juveniles standing on their tails responded differently to different types of information from the environment (Fig. 1). Nonspecific cues such as air movement, in this case associated with moving a syringe tip to within a millimetre of the infective juvenile or applying a puff of air through the syringe, triggered a small increase in jumping. The proportion of individuals jumping increased dramatically when more specific cues indicating the proximity of an insect were introduced. When the tip of a syringe containing larvae of the host *Galleria mellonella* (Lepidoptera) or adults of *Acheta domesticus* (Orthoptera) was introduced to standing infective juveniles, nearly all nematodes were induced to jump.

S. carpocapsae infective juveniles were able to change the direction of jumping in response to information from the environment (Fig. 1). Infective juveniles tended to jump towards the source of the air movement when air movements were small (syringe), but directionality was lost when they were large (puff of air), indicating that large air movements may inhibit the nematode from deriving directional information. When host cues were present, the nematodes tended to jump towards the source of the cues; however, there seem to be two

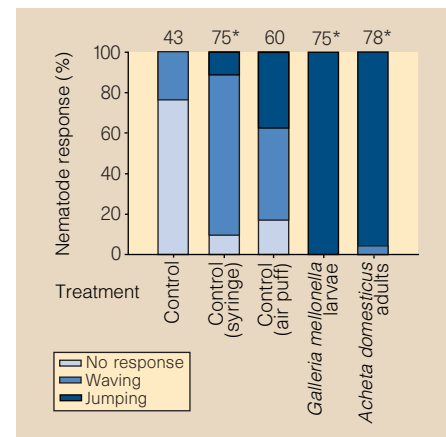


Figure 1 Influence of different cues on the initiation and direction of jumps of *Steinernema carpocapsae* infective juveniles. The following stimuli were presented by moving a Hamilton 10-ml gas-tight syringe to within 1 mm of a standing infective juvenile: control (no syringe); syringe (empty syringe); air puff (syringe was depressed to create a puff of air); or the syringe contained four individuals of either *Galleria mellonella* larvae or *Acheta domesticus* adults. Nematodes either remained in the straight standing posture (no response), waved back and forth while standing (waving), or initiated a jump (jumping). The proportion of individuals jumping towards the cues is shown at the top of each bar; numbers with stars are significantly different in jumping direction (contingency table analysis with chi-square tests at $P < 0.05$). We tested 60 nematodes for each treatment.

separate components to the jumping response because a similar trend was observed in the syringe control. Insect-associated volatile cues are important triggers for jumping behaviour, but it is air movement that influences the direction of the jump.

Directional jumping in response to the close proximity of a potential host may increase the probability of the infective juvenile contacting it. Jumping appears to be part of a range of behavioural traits, including standing and foraging at the soil surface, by which some nematodes are able to exploit ground-dwelling insects as hosts. Understanding the role of jumping in host search behaviour has implications for the successful use of this nematode in the biological control of insect pests⁶.

James F. Campbell*, Harry K. Kaya†

Departments of *Entomology, and †Nematology, University of California, Davis, California 95616, USA
e-mail: jfcampbell@ucdavis.edu

1. Reed, E. M. & Wallace, H. R. *Nature* **206**, 210–211 (1965).
2. Campbell, J. F. & Gaugler, R. *Behaviour* **126**, 155–169 (1993).
3. Maggenti, A. *General Nematology* (Springer, New York, 1981).
4. Kaya, H. K. & Gaugler, R. *Annu. Rev. Entomol.* **38**, 181–206 (1993).
5. Campbell, J. F., Lewis, E., Yoder, F. & Gaugler, R. *Parasitology* **113**, 473–482 (1996).
6. Gaugler, R., Lewis, E. & Stuart, R. J. *Oecologia* **109**, 483–489 (1997).

Non-fertile sperm delay female remating

Sperm competition is thought to be responsible for the tremendous inter- and intraspecific variation in sperm number¹ and size². But why several animals produce a range of sperm types^{3,4}, some of which are incapable of fertilizing the female's eggs⁴, has remained unexplained for nearly 100 years⁵. We have found that non-fertile sperm protect a male's reproductive investment by delaying female remating in the polyandrous green-veined white butterfly, *Pieris napi* (Pieridae).

Lepidoptera have two distinct sperm types: fertile 'eupyrene' sperm and non-fertile 'apyrene' sperm, which lack nuclear material⁶. Apyrene sperm have a distinct developmental pathway, and so cannot be considered aberrant⁶. They are shorter, thinner and contain less mitochondrial material than eupyrene sperm⁶. Both types are transferred to the female at mating, with more than 90% being apyrene sperm⁷, and both migrate to the site of sperm storage, the spermatheca.

The high motility of apyrene sperm suggests that they may aid the transfer of eupyrene sperm, in which case there should be a constant ratio between the two sperm types, whereas males actually vary the

proportion considerably⁷. Alternatively, apyrene sperm may represent nutrients for the female, zygote or eupyrene sperm in storage⁸. This is also unlikely, because nutrient donations in paternally investing species are transferred in a non-ejaculate part of the spermatophore (the sperm-containing packet)⁹. Another possibility is that apyrene sperm may be involved in sperm competition¹⁰. When mating with females who have already been inseminated, apyrene sperm could interfere with rival males' sperm. They may also influence female receptivity, filling the spermatheca and delaying female remating.

We tested this last hypothesis. Females were mated to virgin ($n=22$) or mated males ($n=14$). If apyrene sperm delay remating, we would expect females with more apyrene sperm in storage not to remate. Females were allowed to remate at will for up to ten days after mating (females rarely remate after this time⁹). The two sperm types in the spermatheca originating from the first male were counted, either when the female remated or after ten days if the female did not remate.

We found that female receptivity is related to the number of apyrene sperm in storage. As predicted, remating females have fewer apyrene sperm already present in their spermatheca than females who did not remate, whereas there was no difference in the number of fertilizing sperm stored

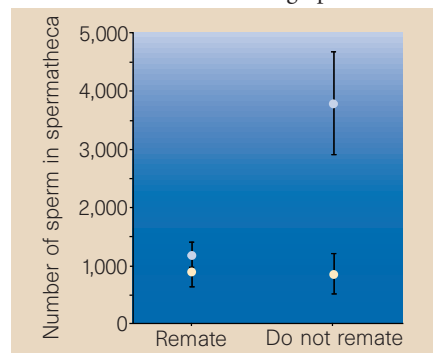


Figure 1 Relation between number of apyrene sperm stored and female remating. Females that do not remate have more apyrene sperm in storage (pale blue circles; log-transformed data, $F_{(1,32)}=5.33$, $P<0.01$). The mating status of her first partner (whether mated or virgin) has no effect on apyrene storage ($F_{(1,32)}=0.05$, $P>0.8$) and there is no interaction between male mating status and female remating ($F_{(1,32)}=0.01$, $P>0.9$). There is no difference in eupyrene sperm number (pale yellow circles; $F_{(1,32)}=1.31$, $P>0.2$), no effect of male mating status ($F_{(1,32)}=2.50$, $P>0.1$) and no interaction ($F_{(1,32)}=0.02$, $P>0.9$). Bars show standard error. There is a positive correlation between days until remating and number of apyrene sperm in the spermatheca (females mated to virgin males, $r=0.64$, $P<0.02$; to mated males, $r=0.63$, $P<0.03$), but no relationship with the number of eupyrene sperm in storage (females mated to virgin males, $r=0.32$, $P>0.3$, $n=13$; to mated males, $r=0.13$, $P>0.7$, $n=12$).

(Fig. 1). Female remating is also related to male mating status. Females receiving smaller spermatophores from mated males (3.6 ± 0.34 (s.e.) mg versus 6.5 ± 0.25 mg; $F_{(1,63)}=48.1$, $P<0.001$) remate more (12 out of 14 versus 13 out of 22, $\chi^2=3.97$, $P<0.05$, 1 d.f.) and sooner (3.4 ± 0.45 versus 5.5 ± 0.55 days, $t=2.95$, $P<0.01$, 24 d.f.) than females receiving larger spermatophores from virgin males. Mated males, despite transferring smaller spermatophores, provide more eupyrene sperm than do virgin males ($8,469 \pm 1,046$ versus $5,757 \pm 911$, $t=2.16$ on log-transformed data, $P<0.05$), as in the related *P. rapae*⁷, but do not transfer more apyrene sperm ($44,954 \pm 5,622$ versus $49,980 \pm 4,475$; $t=0.83$, $P>0.4$, 62 d.f.).

The relation between time before remating and number of apyrene sperm stored indicates that the induction of non-receptivity in females does not depend on a threshold level of apyrene sperm. The quantity of stored apyrene sperm is more variable than the numbers delivered (adjusted coefficient of variation: 111 versus 7). It is not clear whether this variation is due to differences in quality or persistence of apyrene sperm in storage, or in the tendency of females to store apyrene sperm. Nevertheless, controlling for the effect of male mating history, we find that females that store more apyrene sperm do not remate.

Males may be using apyrene sperm to exploit a female system designed to monitor sperm numbers in storage to ensure maximum fertility. Apyrene sperm may be less costly to produce than eupyrene sperm, or more efficient in reducing female receptivity. Instead of containing large amounts of nucleate sperm, males' ejaculates consist mainly of anucleate sperm, delaying female remating and hence reducing the potential for sperm competition.

Penny A. Cook*, Nina Wedell†

*School of Health, Liverpool John Moores University, 79 Tithebarn Street, Liverpool L2 2ER, UK
e-mail: p.a.cook@livjm.ac.uk

†Department of Zoology, University of Stockholm, S-106 91 Stockholm, Sweden and Department of Zoology, University of Melbourne, Parkville, Victoria 3052, Australia

1. Parker, G. A., Ball, M. A., Stockley, P. & Gage, M. J. G. *Proc. R. Soc. Lond. B* **264**, 1793–1802 (1997).
2. Pitnick, S., Spicer, G. S. & Markow, T. A. *Nature* **375**, 109 (1995).
3. Sivinski, J. *Fla. Entomol.* **63**, 99–111 (1980).
4. Snook, R. R. & Karr, T. L. *Curr. Biol.* **8**, 291–294 (1998).
5. Meves, F. *Arch. Mikrosk. Anat. Entomol.* **61**, 1–82 (1902).
6. Friedländer, M. J. *Insect Physiol.* **43**, 1085–1092 (1997).
7. Cook, P. A. & Wedell, N. *Proc. R. Soc. Lond. B* **263**, 1047–1051 (1996).
8. Riemann, J. G. & Gassner, G. *Ann. Entomol. Soc. Am.* **66**, 154–159 (1973).
9. Wiklund, C., Kaitala, A., Lindfors, V. & Abenius, J. *Behav. Ecol. Sociobiol.* **33**, 25–33 (1993).
10. Silberglied, R. E., Shepherd, J. G. & Dickinson, J. L. *Am. Nat.* **123**, 255–265 (1984).

The fashions and foibles of science

Egg and Ego: An Almost True Story of Life in the Biology Lab

by J. M. W. Slack

Springer: 1998. 192 pp. \$24.95, £15.95 (pbk)

Peter A. Lawrence

Jonathan Slack has a good nose for the ridiculous. He is a developmental biologist known for his work on growth factors, and for his excellent text *From Egg to Embryo*. In *Egg and Ego* he casts about more widely, and this book is sometimes straight review, sometimes autobiography, sometimes satirical commentary.

He takes a sardonic look at the lives and aspirations of modern biologists like himself. He picks out and picks on the twisted sense of values, the lost idealism and the many oddities that have become so commonplace in our workaday world. He tells us 'what science is really like' — as distinct from the impersonal and theoretical picture painted by philosophers of science. Any senior biologist will find much in the book to interest him, while a prospective biology student will find out what awaits her (I follow Slack in the choice of genders).

Slack's descriptions of his experiments on growth factors, although lively, are perhaps the least interesting part of the book because they cover old territory. But his descriptions of the increasingly weird practices of the scientific world are new ground. He describes the star system in science, which has got out of hand, with a few well-known people spending a bizarre proportion of their time travelling and talking. Others, who prefer to stay in their labs discovering things, are not so well known because they do not travel, and therefore do not get invited, and therefore remain unknown.

He summarizes how the aspirations of scientists have switched from trying to make illuminating discoveries to a desperate and competitive struggle up a career ladder. He describes how one gets to the 'top'. The top is now defined by the managers who have thrust upon us their own quantitative measures of greatness. In a creative industry like research, where real discoveries are always ahead of their time, these measures are at best crude — like assessing the quality of a songwriter by counting the number of notes he writes in a week. Nevertheless, after a while, what the British political commentator Simon Jenkins has called an "audit society" is created. In such a society, the real purpose of the endeavours becomes forgotten and it devotes itself, not to making the important measurable, but to making the measurable important.

For scientists, chief among these measures is the impact factor of journals. Nowa-



days, as Slack points out, assessment of researchers is not by "the content of the articles, not even by their titles, but just by the names of journals in which they are published". Slack picks the top "fashion journals" in biology as *Cell*, *Nature* and *Science*, which have high impact factors. Yet the impact factor is determined not by the bulk of the papers in the journal, but by a few heavily cited ones. Thus, "most of what the fashion journals contain is actually the same sort of thing the specialist journals carry but dolled up to look a bit special". Slack's papers published in specialized journals were just as good as those published in the fashion journals, both in his opinion and as

measured by the citations they attracted.

He then argues that fashion journals are over-influenced by journalism. He details most entertainingly an experiment in which he submitted a piece of "dilettante-ish nonsense" cooked up as something of a joke which got into *Nature*, in spite of the fact that the reviewers saw through it.

In describing access to the fashion journals, I think Slack has missed a trick, for he has not pointed out that the majority of manuscripts (about 80% of those submitted, in the case of *Nature*) are rejected by the editors without review. Although one can understand the viewpoint of the journals, the outcome of this process is that we have

connived in handing over the captaincy of our fate to the editors of these journals.

These editors are largely full-time professionals who, as Slack writes, are too sensitive to fashion and can be over-suspicious of really new work where the level of rigour cannot be as high as with investigations into familiar territory. The most original papers will be better sent to specialized journals where the editors are more often research scientists, the acceptance rate is higher and papers are more often reviewed. (In *Development*, which Slack names as a leading journal in his field, only 7% of submitted papers are sent back without review.)

But young scientists looking for a career in research may fail if they publish only in specialized journals. They cannot dare to be adventurous, knowing that it is safer to stay near where they started, restricting their project and manicuring their manuscript into whatever form is currently in vogue. Increasingly, many scientists believe that networking with colleagues (possible reviewers of their articles) and with the editors themselves are profitable investments of their time.

Slack ends by telling any would-be biologist to be adventurous, indeed, to fish "in some backwater thought hopelessly unfundable". He notes, "the fashionable stuff always involves lethal competition and is sure to be mined out within a few years". Choose something that interests you personally, "then if you don't get your paper into *Cell* ... you will at least have spent some time doing something ... you felt was really worthwhile". Very good advice, but why have we allowed our career structure to become so distorted that it may prove suicidal to follow it? □

Peter A. Lawrence is in the MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

Darwinizing psychology

The Evolution of Mind

edited by Denise Dellarosa Cummins and Colen Allen

Oxford University Press: 1998. 264 pp.

£25, \$41.13

Andrew Whiten

One of the most significant achievements of evolutionary psychology has been to seriously question the traditional view that we possess a general-purpose intelligence that can analyse any aspect of human experience with equal aplomb. Humanity in the image of a divine being is perhaps the most extreme expression and source of this view. In contrast, an evolutionary perspective reveals human minds, like those of other species, to be imperfect, relatively jerry-built devices



Getting a head

The trebling of brain size in hominid evolution from *Australopithecus* to *Homo sapiens* is explored in this piece of "Brainart" by Evian Gordon, head of the Cognitive Neuroscience Unit at the University of Sydney, Australia. Entitled "Our shared evolutionary history", it comes from *Your Future Self: a Journey*

to the *Frontiers of Molecular Medicine* (Thames & Hudson, \$27.50) by the science popularizer and documentary-maker Hank Whittemore. The book presents images from molecular and cellular biology for non-scientists in an attempt to inspire awe in the unfolding revelation of our inner universe.

that are shaped by natural selection to deal with a specific set of problems in the species' ancestral environment. Minds are expected to have cognitive blind spots.

Recent studies summarized by Gerd Gigerenzer and by Denise Dellarosa Cummins in the two opening chapters of *The Evolution of Mind* show that our celebrated reasoning power works pathetically in some cases. Evolutionary theory can explain this. Gigerenzer's examples are relevant to most scientists who, even if they do not consider themselves divine, take a certain pride in their rationality and numeracy. For example, estimate the probability that a woman with a positive mammogram actually has breast cancer, given that: (1) the probability that a patient has breast cancer is 1%; (2) if the patient has breast cancer, the probability of correct diagnosis from the mammogram is 80%; and (3) if the patient has no cancer the probability of a positive test is 10%. The typical answer given by a large sample of physicians was around 75%. The correct answer is in fact one order of magnitude smaller.

Gigerenzer's explanation is that ancestral reasoning processes never received input in the form of such probabilities and so are not naturally adapted to interpret them. By contrast, a common ancestral form of information was the raw frequencies of different events. When the cancer problem is cast in simple frequencies, people are more likely to arrive at the correct answer. Imagine

answering the question again, having seen 10 cases out of 1,000 with both a positive mammogram and cancer, and 100 cases with a positive mammogram and no cancer.

Gigerenzer's beautifully executed opening chapter is exciting for at least two reasons. The first is the revolutionary implications for our view of the human mind: human reasoning may not be homogeneous but may run on both rational and non-rational tracks that only make sense from the perspective of evolutionary psychology. In Gigerenzer's own terms, we are moving towards an understanding of our "bounded rationality". The second reason for excitement is that these insights not only are relevant to academic disciplines, but also have serious practical implications, ranging from medicine to the legal system to the teaching of statistical reasoning.

The Evolution of Mind is therefore a timely collection. The authors of its 10 chapters are drawn from departments of psychology, philosophy, biology and anthropology, and the book covers both human and non-human minds. The substantial chapter by the primate ethologist Marc Hauser and the developmental psychologist Susan Carey is particularly important in this respect because it counters the dearth of truly comparative studies. They apply the same experimental techniques in probing the minds of both monkeys and pre-verbal human infants, to discriminate shared cognitive

primitive concepts (such as basic quantifiers as one/another) from mental operations emerging in only one of these species.

However, the book does not attempt to represent evolutionary psychology as a whole. There is no coverage of the burgeoning literature on mate-choice or altruism, for example. 'Mind' is interpreted as the 'higher' cognitive functions, with two chapters devoted to reasoning, two to numeracy and two to language. Even here, there is no attempt to be comprehensive; cognitive mapping, tool use or cultural learning, for example, are not included. Since the rationale for this selectivity is not explained, the book has the feel of being a whimsical sampling of current thought on the evolution of mind. Thus, although the well-written chapters could be read individually as introductions to their specific areas, the book is not an overview for students. That said, some contributions are original and merit the attention of advanced scholars in both evolutionary psychology and related disciplines.

Some of the philosophers' contributions are unusually refreshing, particularly since they engage the concerns of practising scientists, as in Elliot Sober's chapter on the appropriate interpretation of Lloyd Morgan's famous canon: "In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale."

Two chapters deal with the underlying rationale of evolutionary psychology more directly. Colin Allen and Eric Saidel build the case that although a certain cognitive structure may be uniquely human, the origins of at least some of its components may well be evolutionarily traceable. They illustrate this by examining the primate counterpart of language's ability to refer a listener to topics.

In the final chapter, Lawrence Shapiro identifies the core of evolutionary psychology: "what Darwin did for psychology is to license and ground the ascription of teleological functions to mental processes." Readers new to, or suspicious of, evolutionary psychology should perhaps read this chapter first. □

Andrew Whiten is in the School of Psychology, University of St Andrews, Fife KY16 9JU, UK.

More on the mind

Reaching into Thought: The Minds of the Great Apes

edited by Anne E. Russon, Kim A. Bard & Sue Taylor Parker

Cambridge University Press, £24.95

The Evolution of Consciousness

by Euan M. Macphail

Oxford University Press, £15.99, \$28

How the Mind Works

by Steven Pinker

Penguin, £9.99

Earth's internally generated motions

Theoretical Global Seismology

by F. A. Dahlen and Jeroen Tromp

Princeton University Press: 1998. 1,025 pp.

\$80, £60

Paul G. Richards

Global seismology began in the 1880s when the first teleseismic recording was made in Europe of an earthquake in Japan. For decades the subject developed as the study of the seismic body waves that propagate through Earth's deep interior, and the surface waves that are confined to the crust and upper mantle. Such waves can now be faithfully recorded across vast ranges of frequency, wavelength, distance and amplitude; and we study them for what they can tell us about the seismic source — perhaps an earthquake or an explosion — and about the Earth's structure through which they have travelled.

Since the underlying seismic motion can be quantified using linear wave equations, and since the Earth is a body of finite size, we know that all seismic motions of our planet can be expressed as a sum of the Earth's normal modes. Conceptually, every seismogram of every earthquake and explosion is a sum of the same set of decaying sinusoids, each of which begins its oscillation at every point of the planet when the earthquake begins. In practice, the set sums to zero until the first body wave arrives, and thereafter it is a representation of the ground motion at that recording station.



Communing with nature: a 1917 seismologist listens to the rumbling of Italy's Solfatara volcano.

Theoretical Global Seismology by F. A. Dahlen and Jeroen Tromp consists of more than 1,000 pages and 10,000 equations and is motivated by the simplicity and elegance of this idea of mode summation. It is the first book to pull together the necessary details to address the complications arising from the fact that our planet departs slightly from elastic behaviour and is far from homogeneous.

First, the authors present the continuum mechanics that are needed to quantify features exhibited by the Earth — anelasticity, heterogeneity, anisotropy, high internal stress, rotation and spontaneous failures of

New in paperback

At the Fringes of Science

by Michael W. Friedlander

Westview Press, £11.95, \$17

The Call of Distant Mammoths:

Why the Ice Age Mammals Disappeared

by Peter D. Ward

Springer, \$15, £11.50

Colonialism and its Force of Knowledge: The British India

by Bernard S. Cohn

Princeton University Press, \$17.95, £12.95

The Astronomer's Universe

by Herbert Friedman

W. W. Norton, \$14.95, £10.95

Sir Aurel Stein: Archaeological Explorer

by Jeannette Mirsky

University of Chicago Press, \$20, £15.95

Redeeming Culture: American Religion in an Age of Science

by James Gilbert

University of Chicago Press, \$19, £15.25

The Kindly Dr. Guillotin: And Other Essays on Science and Life

by Harold J. Morowitz

Counterpoint, £8.95, \$12.50

Linear Programming and Extensions

by George B. Dantzig

Princeton University Press, \$29.95, £23.95

From Savage to Negro: Anthropology and the Construction of Race, 1896–1954

by Lee D. Baker

University of California Press, \$17.95, £12.95

Fairweather Eden

by Michael Pitts & Mark Roberts

Arrow, £8.99

Hooke's law (otherwise known as earthquakes). Then they set out methods for quantifying the free oscillations of a spherically symmetrical Earth model, and show how these oscillations can be classified into groups that are affected by particular regions of the Earth's interior. Finally, they show how ellipticity and other forms of lateral variability can be handled.

Although the book's main emphasis is on normal modes, asymptotic methods are given to explain the principal features of body and surface waves. Most applications of seismology, including many of our discoveries about the Earth's internal structure, have been based on such asymptotic methods.

The book's presentation is authoritative, and graduate students and their supervisors for years to come will turn with gratitude to *Theoretical Global Seismology* to learn how to handle such complications as the effect of high initial deviatoric stress on the body force equivalent of an earthquake, or to interpret an accelerometer record at low frequencies when there can be significant contributions from changes in elevation (free air) and the overall gravity field.

Dahlen and Tromp's book shows us that seismology is a highly specialized science in that it focuses just on the internally generated motions of our planet. But, on the other hand, because we can study the Earth year after year and century after century, seismology also forces us to generalize our understanding of linear wave propagation. This we do in order to quantify the effects of inhomogeneity, high initial internal stress and discontinuities across a fault. Our new understanding is then available for other fields, for example waves in layered semiconductors.

The theoretical seismology in this book is not merely an intellectual achievement of a very high order, it is driven by the need to understand our planet and will be the basis of applied seismology and other practical studies of wave generation and propagation for decades to come. □

Paul G. Richards is at the Lamont-Doherty Earth Observatory, Palisades, New York 10964, USA.

Science in culture

π -fetishism

π : A Darren Aronofsky film
Artisan Entertainment: 1998
 Graham Farmelo

π is, I believe, the first film to be named after a mathematical symbol. This is a sign of the times: mathematics is seeping into popular culture, notably through books that have made Andrew Wiles's proof of Fermat's last theorem and the life of the number theorist Paul Erdős the stuff of recreational chatter.

No sooner have the educational Jeremiahs convinced us that mathematical standards in schools are both low and falling, at least in the West, than the subject becomes a big-business leisure activity among people who cheerfully admit they cannot cope with long division.

Hollywood discovered the new appeal of mathematics two years ago in the box-office hit *Good Will Hunting*, director Gus Van Sant's engaging tale of a delinquent mathematical genius. Set in Boston, the Athens of America, this was a surprisingly unsentimental portrayal of academic life in which mathematicians were for once played as more-or-less normal people and not as unkempt sociopaths with foreign accents as thick as treacle.

π is more challenging fare, cinematically and mathematically. Like *Good Will Hunting*, it features a mathematician, but is — superficially at least — more concerned with how mathematics might be applied to the world. However, π is not another crowd-pleasing blockbuster from the dream factory, but an accessible art-house film, produced from the independent sector. It first shot to notoriety after winning a Directing Award at the 1998 Sundance Film Festival.

Director and co-writer Darren Aronofsky sets his film in some of the more unsavoury parts of New York. The action revolves around the misanthropic number theorist Max Cohen (played by Sean Gulleette), a certifiable π -fetishist and, of course, a genius (why do film-makers seem to think that all mathematicians are brilliant?). He works alone at home in his seedy little apartment, an Alfred Brendel of the computer keyboard,

π -fetishist Max Cohen (Sean Gulleette).

looking as if he's recovering from a lobotomy.

Cohen is convinced that the (almost random) digit sequence of π holds the key to understanding the pattern of daily changes in the New York stock market, which he monitors obsessively during his forays into the real world. "Mathematics is the language of nature," the narrator repeatedly intones, but Cohen does not bind himself to the rigours of scientific analysis, and so, as his mentor warns him, he is in danger of practising mere numerology.

As Max closes in on his solution, he is pursued by Wall Street hustlers intent on financial domination and by a Kabbalah sect trying to unlock the secrets behind their ancient holy texts. So we are in strange territory here, in the nether world of impure mathematics, tenuously connected to the real world via a 24-carat eccentric and a host of gargoyles. You can't help feeling sorry for Cohen's neighbour: would you like to live next to someone whose work involves washing a human brain down his bathroom sink?

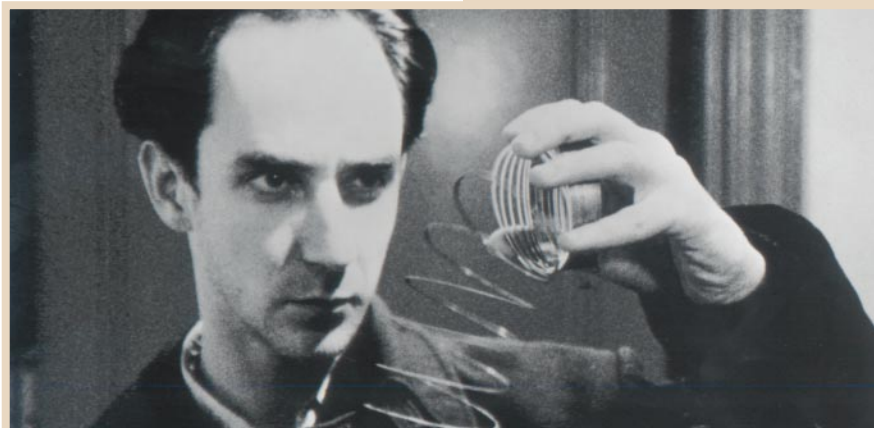
Aronofsky does not aim to give us a mathematics lesson, so he does not consider in any detail what has made the number π such a rewarding mystery for mathematicians over the past 4,000 years. But he does gently try to interest us in the possibility of understanding the patterns we see around us. In one scene, Cohen looks out across a beach and we are invited to consider with him what might underlie the pattern of waves on the shore, of sea shells and of the light shimmering on the ocean.

Such quiescent moments are rare; π is in the main a sharply cut thriller, suffused with a bleakness intensified by grainy black-and-white photography and the menacing electronic soundtrack. The script is taut, the production inventive and the direction striking. It all amounts to the most imaginative debut for a writer-director since Quentin Tarantino gave us *Reservoir Dogs*.

But π is not really about mathematics. Like Bertolt Brecht's *Life of Galileo*, which is less concerned with science than with the inability of a scientist to cope with the moral consequences of his work, Aronofsky's film uses his theme of mathematics-inspired hokum as a backdrop for something different — an off-beat analysis of paranoia.

Now that mathematics is 'cool', its ideas have new currency in the arts. For me, this is both welcome and surprising. When I was an undergraduate, I attended an entertaining lecture on *Principia Mathematica*, Bertrand Russell and Alfred North Whitehead's mind-stretching account of the foundations of mathematics. But if Russell and Whitehead were so smart, our professor mused, why didn't they have the wit to sell the film rights? In 1973, that remark brought the house down. It wouldn't surprise me now if the laugh turns out to be on us. □

Graham Farmelo is at the Science Museum, South Kensington, London SW7 2DD, UK.



Changing sources of nutrients during four million years of ecosystem development

O. A. Chadwick, L. A. Derry, P. M. Vitousek, B. J. Huebert & L. O. Hedin

As soils develop in humid environments, rock-derived elements are gradually lost, and under constant conditions it seems that ecosystems should reach a state of profound and irreversible nutrient depletion. We show here that inputs of elements from the atmosphere can sustain the productivity of Hawaiian rainforests on highly weathered soils. Cations are supplied in marine aerosols and phosphorus is deposited in dust from central Asia, which is over 6,000 km away.

Terrestrial biogeochemists traditionally distinguish atmospherically derived from rock-derived elements^{1,2}. Atmospherically derived elements such as carbon and nitrogen have an important gas phase; they enter terrestrial ecosystems either through biological processes such as photosynthesis and biological N₂ fixation, or through deposition from the atmosphere (for example, dissolved in precipitation or by dry deposition of particles and gases). Rock-derived elements such as calcium, magnesium, potassium and phosphorus are important constituents of minerals; they enter terrestrial ecosystems through partial or complete dissolution of these minerals (chemical weathering). A central conceptual model for the development of soils and ecosystems is built around this distinction³; new substrates that are laid down by volcanic eruptions, glacial recessions, or other processes that initiate the formation of wholly new soils and ecosystems, generally lack atmospherically derived elements, particularly nitrogen and carbon, but they are rich in minerals that can supply rock-derived elements. Consequently, rates of plant production and other ecosystem processes are often constrained by the supply of nitrogen in young developing ecosystems, or such systems are dominated by plants with well developed N₂-fixing symbioses, such as alder and various legumes^{4–6}.

This model assumes that, over time, the stock of weatherable minerals in soil is depleted and rock-derived elements are lost without replacement, whereas atmospherically derived elements

can be replenished continuously. Eventually, in the absence of disturbance of the substrate, ecosystems can reach a terminal steady state of profound and irreversible depletion of rock-derived elements. Phosphorus in particular is commonly implicated as the master regulator of ecosystems in the long term^{3,7}, although other rock-derived elements (Ca, Mg, K) are more mobile than phosphorus and should be depleted more rapidly.

Several lines of evidence suggest that the sharpness of the dichotomy between rock- and atmospherically derived elements has been overstated, with significant consequences for our understanding of biogeochemistry. These include: (1) studies of the mass balance of elements in watersheds show that the quantities of putatively rock-derived elements dissolved in precipitation can be a substantial fraction of the weathering sources of those elements^{8,9}; (2) large quantities of dust are entrained in the atmosphere in arid regions, and transported globally^{10–14}; and (3) atmospheric deposition of all elements may be understated by standard measurements. The dry deposition of particles and gases to vegetation surfaces can be substantial^{15,16}, and where cloudwater is deposited in forests it is likely to be a particularly important source of elements^{17,18}.

We evaluated sources of biologically significant elements and their implications for ecosystem functioning across a developmental sequence of sites in the Hawaiian islands. Although recent analyses of ecosystem dynamics have focused on climate and climate change, in Hawaii as elsewhere^{19–21}, the quantity and availability of nutrient elements in the soil also control plant productivity and other aspects of ecosystem functioning²². Moreover, terrestrial ecosystems are altered by anthropogenic changes in nutrient supply^{23,24} and by interactions between climate change and nutrient availability²⁵ at least as much as by climate change itself.

Hawaii is a model ecosystem

The Hawaiian islands are formed by a stationary convective plume (hot spot) that taps lava from the upper mantle. The volcanoes grow for about 600,000 yr, drifting to the northwest on the Pacific lithospheric plate²⁶. Each island (and each volcano within an island) is progressively older along a transect from active volcanoes in the southeast to the oldest islands in the northwest²⁷. Along the island chain, we located six sites that differ markedly in the age of their underlying substrate, from 300 through to 2,100, 20,000, 150,000, and 1,400,000 to 4,100,000 years old (Fig. 1). The sites are similar in their current climate, with a mean annual temperature of 16 °C, and about 2,500 mm of precipitation annually. The substrate is basaltic rock admixed with tephra and pumice, with an initial chemistry that varies relatively little in space or time²⁸. Over time, the topography evolves from classical shield volcanoes with gentle slopes, to deeply dissected landforms²⁹; however, we were able to

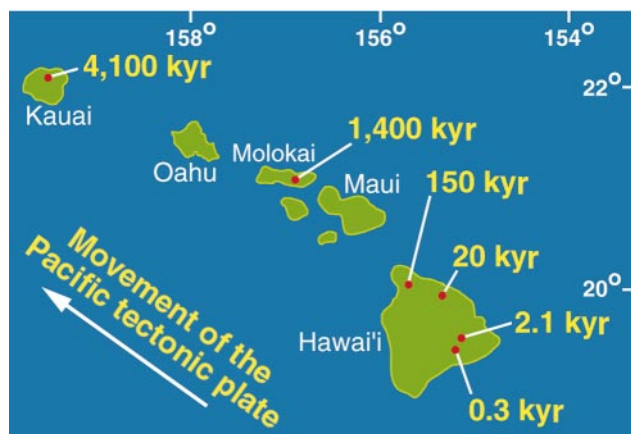


Figure 1 Location of the study sites. The youngest two sites are on the still-active Kilauea volcano; the 20,000-yr-old site is on Mauna Kea, the 150,000-yr-old site is on Kohala, the 1,400,000-yr-old site is on East Molokai, and the 4,100,000-yr-old site on Kauai.

locate residual shield surfaces on even the oldest sites. None of the sites has been cleared or systematically altered by people.

Finally, the Hawaiian islands form the most remote archipelago on Earth. Relatively few species reached Hawaii naturally (blown there in storms or attached to migrating birds) in the tens of millions of years before human discovery³⁰. Some of these species then radiated to occupy a range of environments far broader than do continental species, resulting in an astonishing degree of biological similarity among sites. Overall, this near constancy in geology, topography, climate and biota, coupled with a very wide and well characterized range in substrate age, makes the Hawaiian islands an extraordinary resource for studies of soil and ecosystem development.

We cannot analyse soil and ecosystem development by following a single site through time; these processes are too slow for that. Rather, we are comparing sites with substrates of different ages: in so doing, we implicitly assume that this substitution of space for time^{31,32} will provide an insight into factors that control ecosystem development on geological timescales. We do not assume that all sites have been influenced by identical conditions throughout their history; older sites have been subjected to greater climatic variation^{33–35} and isostatically and thermally driven subsidence^{26,36}. Nevertheless, along the transect the constant substrate has been exposed to the effects of environmental and biological forcing for substantially different lengths of time.

Ecological changes and substrate age

We evaluated patterns in soils, nutrient availability, forest composition and structure, net primary production (NPP), and nutrient limitation to NPP across the developmental sequence. The soils of the two youngest sites are little weathered; their properties are strongly influenced by the properties of the coarse-textured pumice-rich parent material. The two intermediate-aged sites

have highly weathered andisols that contain high levels of chemically active non-crystalline minerals that readily complex organic matter and adsorb phosphorus, whereas the soil at the 1.4-million-year-old (Myr) site is an ultisol in which most of the chemically reactive non-crystalline minerals have transformed to less-active clays, with loss of much of the nutrient-holding capacity of the younger soils. Finally, the soil at the oldest site is an oxisol that has lost most of its active minerals to weathering and leaching; most of the remaining soil is made up of inert iron and aluminium minerals typical of highly weathered tropical soils, with no capacity to supply nutrients and little ability to retain recycled or added nutrients³⁷.

The biological availability of essential plant nutrients also varies systematically along the sequence; the young sites have high concentrations of available cations (Ca, Mg, K) and little available nitrogen and phosphorus; the intermediate-aged sites have fewer available cations and more nitrogen and phosphorus; and the oldest site has low phosphorus and cation availability and high nitrogen availability (Fig. 2a, b)^{38,39}. Low nitrogen availability in the young sites reflects a lack of nitrogen in the substrate and the time required to accumulate nitrogen from the atmosphere⁴⁰. The mechanisms that control patterns in available cations and phosphorus are discussed below.

The species composition of the forest vegetation is remarkably consistent across the sites, despite the wide range in soils and available nutrients. The native tree *Metrosideros polymorpha* makes up 80–88 per cent of each forest's basal area, and several other species are found in all of the sites⁴¹. We analysed element concentrations in leaves of *Metrosideros* and other species across the sequence to provide an assessment of nutrient availability as it is experienced by vegetation⁴². The leaves reflect soil nutrient availability quite closely, with peak cation concentrations in young sites and peak concentrations of nitrogen and phosphorus in the intermediate-aged sites (Fig. 2c, d); the main difference between soil and

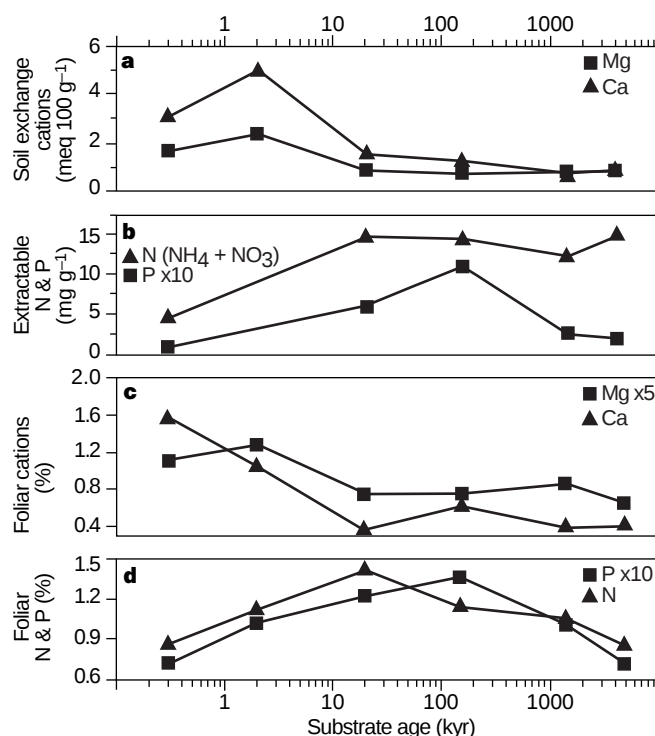


Figure 2 Plant-available nutrients in soil and nutrients in leaves. **a**, Ca and Mg exchangeably bound to soil colloids³⁶. **b**, Phosphate-P and ammonium-N plus nitrate-N in soil solution or exchangeably bound to colloids, forms that are readily available for plant uptake³⁹ (P. A. Matson, unpublished data). **c**, Ca and Mg in *Metrosideros* leaves⁴². **d**, P and N in *Metrosideros* leaves⁴².

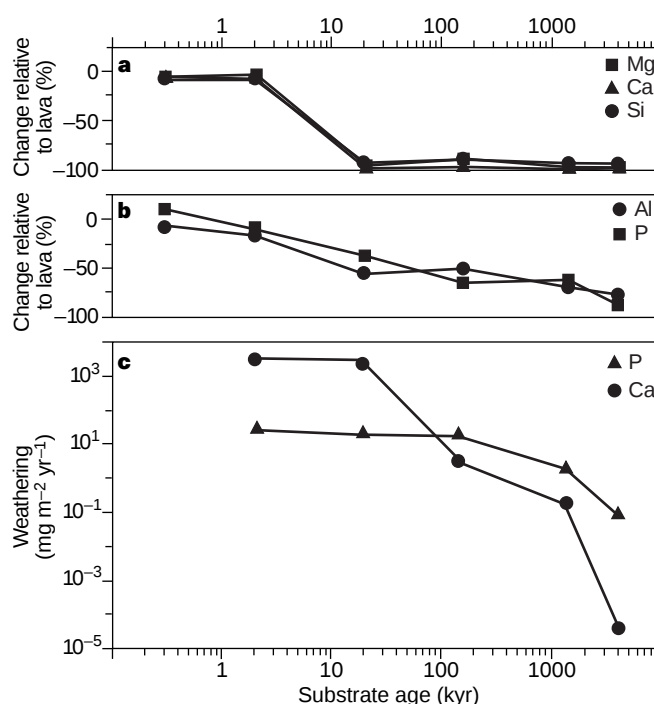


Figure 3 Change in soil element content integrated over the top metre of soil, compared with element contents in the lava parent material³⁸. Measured element concentrations (**a**, Mg, Ca, Si; **b**, Al, P) were corrected for changes in density and loss of mass during soil formation⁷¹. In **c**, the rates of loss of P and Ca are calculated using the mass of an element lost between two sites and the corresponding difference in age.

plant measurements is that the nitrogen concentration in leaves declines (together with phosphorus) in the oldest site, whereas soil availability of nitrogen remains high. When the leaves rich in nitrogen and phosphorus in the intermediate-aged sites senesce and fall, they decompose and release nutrients much more rapidly than do decomposing leaves in either young or old sites—providing a positive feedback to high nitrogen and phosphorus availability in the intermediate-aged sites³⁹.

The net primary production of the forests follows a pattern similar to that of nitrogen and phosphorus availability, peaking in the relatively fertile intermediate-aged sites^{38,43}. We evaluated nutrient limitation to forest growth/NPP in three sites on the sequence using factorial fertilization experiments, and found that nitrogen (and no other element) limits forest growth in the youngest site, whereas phosphorus (and no other element) does so in the oldest site^{43–45}. Adding nitrogen and phosphorus together (but not for either alone) stimulated growth in the 20,000-yr-old intermediate-aged site⁴⁰. Additions of other important rock-derived cations (Ca, Mg, K) and micronutrients (Cu, Fe, Mn, Mo, Zn) had no significant effect on plant production in any of the sites.

Sources of nutrients

Overall, patterns in soil and ecosystem properties along the Hawaiian sequence are consistent with the model that young ecosystems are characterized by low concentrations of atmospherically derived nutrients, whereas old ones are characterized by low concentrations of rock-derived nutrients³. Moreover, the fertilization experiments provide the first clear experimental test of the model's applicability to nutrient limitation in unmanaged ecosystems⁴⁰. However, these results also raise a number of questions—most notably, why does low phosphorus availability limit NPP in ecosystems on old sub-

strates, when the rock-derived cations are much more readily leached from ecosystems than phosphorus is?

We answered this question by measuring both weathering and atmospheric sources of elements along the Hawaiian sequence. The constancy of these sites in climate, geology, topography and species composition provides a consistent background against which the sources of elements can be determined; the range in the age of the sites allows a direct calculation of weathering inputs. We determined the rock-derived elements remaining in primary mineral forms and those retained within the soil as a whole, in each of the sites along the sequence. Initially, rock-derived elements are weathered rapidly, followed by declines to very low weathering rates in the oldest sites. The loss of elements to leaching leads to soil collapse. In the older sites, most of the mass of rock present at the initiation of ecosystem development has been lost; a metre of highly altered soil is all that remains of more than 10 m of rock³⁸.

Most of the calcium, magnesium and silicon weathered from primary minerals was lost rapidly; in contrast, phosphorus and aluminium were retained within the soil for much longer after their release (Fig. 3a, b). The mobility of these elements differs because calcium and magnesium are large divalent ions that are relatively weakly sorbed to soil colloids, and silicon in the undissociated form of $\text{Si}(\text{OH})_4$ is not strongly attracted to soil colloids. In contrast, hydrated aluminium and phosphorus compounds form dissociated weak acids that are highly reactive with mineral and organic surfaces and hence are much less susceptible to leaching. Consequently, weathering-derived calcium is more available to plants early but leaches much more rapidly than phosphorus (Fig. 3c). Still, despite its mobility and the lack of a weathering source of calcium in the older sites, the fertilizer experiments show that calcium does not limit ecosystem productivity anywhere⁴⁰. These observations suggest

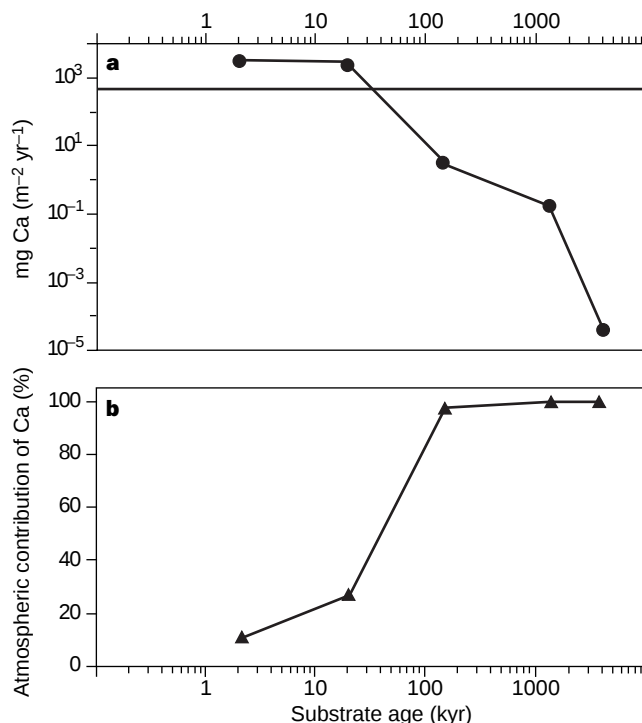


Figure 4 Comparison of Ca inputs from substrate and the atmosphere. **a**, Rate of Ca loss from substrate (Fig. 3c) plotted with the present addition of Ca in rain water and cloud water. **b**, The per cent of total Ca inputs that are derived from the atmosphere, calculated by dividing the atmospheric contribution by the sum of atmospheric and weathering (actually loss from substrate) contribution.

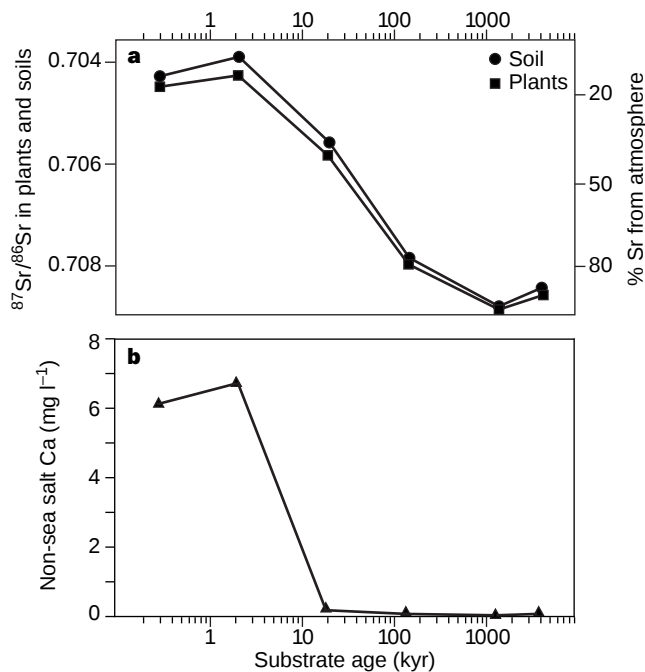


Figure 5 Indicators of the sources of Sr and Ca measured for each site. **a**, The $^{87}\text{Sr}/^{86}\text{Sr}$ ratio of Sr extracted from soil exchange sites and from *Metrosideros* leaves. The isotope ratio is primarily constrained by the value for Hawaiian lava of 0.7035 and a value for sea water (and hence rainfall in Hawaii) of 0.7092 (ref. 48). **b**, The concentration of Ca in soil solution collected by lysimeters inserted below the primary rooting zone in the soils. The measured values have been normalized using a standard sea-salt ratio to reflect Ca released to soil solution by mineral weathering only.

that there must be an atmospheric source of calcium that maintains calcium pools and plant productivity in the older sites.

Atmospheric inputs of calcium were measured directly in the 300-yr-old site by independent analysis of the quantity and chemistry of rainfall, dry deposition, and cloudwater interception¹⁸. We found that the rate of input of calcium, mostly from marine aerosols, is $600 \pm 400 \text{ mg Ca m}^{-2} \text{ yr}^{-1}$. These atmospheric inputs are low relative to inputs from rock weathering in sites of less than 20,000 yr old, but as weathering inputs decline the atmosphere becomes the dominant source of calcium in sites older than 20,000 yr (Fig. 4a, b).

We tested the magnitude and implications of atmospheric additions of calcium by using strontium isotopes. The $^{87}\text{Sr}/^{86}\text{Sr}$ ratio of a sample and its isotopically distinct sources has been used to evaluate atmospheric compared with rock sources of strontium, and by implication, of its fellow alkaline earth elements calcium and magnesium⁴⁶. However, differential weathering rates of minerals in rocks often leads to long-term fractionation of weathering sources and their isotopic values, and inhomogeneities in atmospheric sources have also made such calculations difficult to constrain^{47,48}. Fortunately, Hawaiian lavas and their minerals have a relatively homogeneous chemical and isotope composition which is distinct from that of continental crust and sea water. We demonstrated that strontium in plants and soils shifts from being over 90 per cent rock-derived at the 300- and 2,100-yr sites, to a mixture of rock and atmospheric sources at the 20,000-yr site, to over 80 per cent atmospherically derived by 150,000 yr and thereafter (Fig. 5a)⁴⁸. These results can be extended more or less directly to calcium and magnesium; potassium is even more mobile than the alkaline earth elements in this system, and it should become dominated by atmospheric sources sooner. As an independent check on the importance of atmospheric sources, we measured the concentrations of the principal non-seasalt-derived (that is, rock-derived) cations in soil solution below the rooting zone of the sites. Rock-derived calcium contributes substantially to solution chemistry in young sites, but it disappears entirely by the 150,000-yr-old site (Fig. 5b). Overall, it is clear that the atmosphere becomes the dominant source of most putatively rock-derived elements in older sites. The dilute contribution of sea-salt cations in rainfall and intercepted cloudwater is enough to maintain ecosystem productivity after the rock sources are depleted.

The dynamics of phosphorus differ from those of the other rock-derived elements in that, even after its release by weathering from

primary minerals, its low mobility through soil means that the rock-derived element remains important into much older sites. However, most phosphorus has been lost from the oldest sites (Fig. 3b), and that remaining is largely in forms that are unavailable or only slowly available to plants³⁹. Over time, plant production becomes constrained by the low biological availability of phosphorus.

Could the atmosphere make any significant contribution to phosphorus supply in these older systems⁴⁹? The seawater contribution of phosphorus to marine aerosols is vanishingly small, because uptake by marine organisms maintains very small amounts of phosphorus in surface waters⁵⁰. The largest source of atmospheric phosphorus must be transported by dust through the troposphere on the prevailing westerly winds from central Asia (Fig. 6)^{51–54}. Rates of dust deposition are now relatively small but were several-fold greater during the last glacial period when Asia was less vegetated and overall wind speeds were faster⁵².

The presence of exotic minerals deposited by aeolian processes in Hawaiian soils is well established^{55,56}. Even averaged over time, Asian dust could make only a small contribution to inputs of Ca, Mg or K in comparison to marine sources, but its relative contribution could be substantially greater for phosphorus⁴⁸. Over the past 2–3 Myr, the rate of dust deposition in the Pacific ocean near Hawaii is estimated at $250\text{--}500 \text{ mg m}^{-2} \text{ yr}^{-1}$ (ref. 52). If dust delivers an average crustal phosphorus concentration of 700 p.p.m. (ref. 57), then about $0.17\text{--}0.35 \text{ mg P m}^{-2} \text{ yr}^{-1}$ would be added to the soil. However, dust can be scavenged by rainfall and probably cloudwater, and, owing to orographic processes, the sites receive much more rainfall than the open ocean. Can we estimate phosphorus inputs from the dust deposition to our sites?

Just as the isotope composition of strontium can be used to distinguish the source of alkaline earth elements, refractory trace elements and isotopes of neodymium, a rare earth element, may be useful for resolving the contribution of phosphorus from continental dust. Phosphate minerals such as apatite also carry relatively high concentrations of many refractory trace elements, including rare earths and thorium, making these elements useful tracers of phosphorus. Differences between the $^{143}\text{Nd}/^{144}\text{Nd}$ ratios (or other tracers) of Asian dust and Hawaiian rock can be used as endmembers to calculate the amount of exogenous material that has been added to the soil as dust. If we then know the P/Nd ratio of the endmembers, we can calculate the amount of phosphorus contributed from the atmospheric source. Hawaii is an ideal site to use geochemical tracers in this manner because there is a strong contrast between the chemical properties of continental dust and Hawaiian basalts^{58–61}. We applied this approach using $^{143}\text{Nd}/^{144}\text{Nd}$, Eu/Eu* and Hf/Th as tracers (see Box 1) and calculate an input value of $0.9 \pm 0.3 \text{ mg P m}^{-2} \text{ yr}^{-1}$ added by dust, which is significantly more than the phosphorus contributed to the open ocean.

In contrast to calcium, which is lost from these ecosystems shortly after being weathered from primary minerals, rock-derived phosphorus continues to contribute to ecosystems for more than 1 million years (Fig. 8a). However, losses eventually accumulate to the point where most lava-derived phosphorus has been leached. By the oldest site on the sequence, phosphorus provided by Asian dust is a substantially larger source than is the parent rock (Fig. 8a, b). Plant production in this old site now is limited by phosphorus. However, the limitation would probably be far more severe were it not for phosphorus transported more than 6,000 km from central Asia, most of it over 10,000 years ago.

Broad implications of the Hawaiian model

We can draw several conclusions from our research on the sources of nutrients to Hawaiian ecosystems. First, the distinction between rock-derived and atmospherically derived elements is not sharp; on this sequence of sites, the atmosphere becomes the dominant source of all major biological nutrients apart from phosphorus in less than 100,000 years. It is input from the atmosphere that sustains

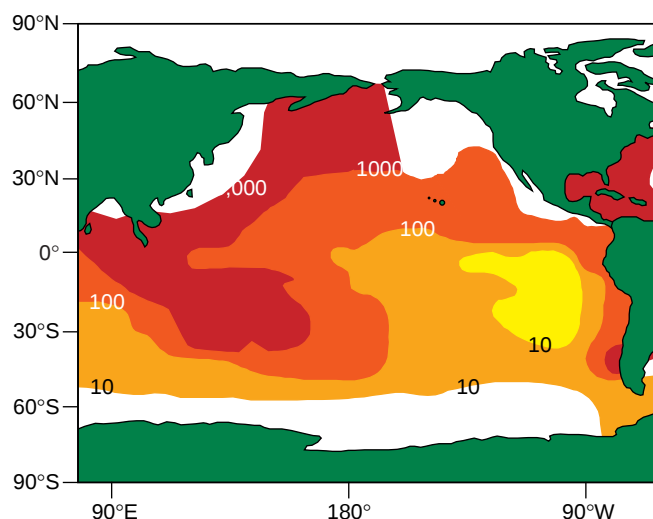


Figure 6 Map showing estimates of the long-term (integrated glacial plus interglacial) rate of dust deposition to the Pacific Ocean (from ref. 58). The isopachs ($\text{mg m}^{-2} \text{ yr}^{-1}$) are based on models of atmospheric dust transport⁵⁸ and are in general agreement with data collected from numerous ocean cores⁵².

Box 1 Geochemical tracers of atmospheric inputs to soils and ecosystems.

To quantify the deposition of the phosphorus by dust, we use the fact that soils accumulate relatively immobile trace elements^{67,68}. Differences in the isotopic and trace-element composition of Asian dust compared with Hawaiian basalt and its weathering products allow us to define endmembers for contributions from dust relative to basalt^{69–61}. We use quantities and mixing ratios of these elements to estimate inputs from dust relative to basalt in the 150,000-yr soil profile, and then extend the analysis to determine phosphorus deposition by dust.

The chemical characteristics of Hawaiian lavas differ substantially from those of average continental crust, and allow us to define unique and useful endmembers on the basis of isotopic and trace element data, including the rare earth elements (REE). Two useful measures of REE sources that are not greatly affected by soil processes are the Eu anomaly, defined as $\text{Eu}/\text{Eu}^* = [\text{Eu}]/([\text{Sm}] + [\text{Gd}])$ (the ratio between the 'expected' Eu concentration as interpolated between Sm and Gd, and the observed value), and the $^{143}\text{Nd}/^{144}\text{Nd}$ ratio, expressed as ϵ_{Nd} (ref. 69). Fresh Hawaiian basalts at our sites have a slight positive Eu anomaly (Eu/Eu^* is 1.0–1.1), and $\epsilon_{\text{Nd}} = +7$, whereas eolian sediments from the Pacific Ocean near Hawaii have a pronounced negative Eu anomaly typical of continental materials ($\text{Eu}/\text{Eu}^* = 0.6$), and $\epsilon_{\text{Nd}} = -10$ (ref. 58) (Table 1). REE patterns in soils from our study sites show a range of patterns that are intermediate between these endmembers. Using mixing equations, we can calculate the contribution from each source.

Similarly, ratios of refractory trace elements can also be used to distinguish continental and basalt sources. In particular, Th and Hf exhibit low mobility in the soil environment. Th/Hf ratios in Hawaiian basalts are near 0.25, whereas those in continental crust are near 1.8 (Table 1). Data from Hawaiian soils suggest that Th/Hf is not significantly changed by weathering processes, providing another tracer for determining the contribution of continental dust to Hawaiian soils.

The fraction of the Nd in a soil sample that is derived from dust can be calculated from the isotopic mass balance (a similar calculation can be used

for Eu/Eu^* or Th/Hf):

$$f_{\text{Nd}}^{\text{dust}} = \frac{\left(\frac{\epsilon_{\text{Nd}}^{\text{soil}} - \epsilon_{\text{Nd}}^{\text{basalt}}}{\epsilon_{\text{Nd}}^{\text{dust}} - \epsilon_{\text{Nd}}^{\text{basalt}}} \right)}$$

Individual soil horizons in Hawaii have $\text{Eu}/\text{Eu}^* = 0.63$ and ϵ_{Nd} as low as -7 , and Th/Hf as high as 1.5, indicating that $>80\%$ of the Nd, Eu or Th can be derived from Asian sources (Fig. 7).

We integrate the dust Nd (or Eu or Th/Hf) contribution to the soil profile using the soil Nd concentration $[\text{Nd}]$, the mass fraction of Nd derived from dust sources ($f_{\text{Nd}}^{\text{dust}}$) and the dry bulk density (ρ) as functions of depth (z) to obtain the integrated flux of Nd that the soil has received over time:

$$\text{Nd}_{\text{dust}}^{\text{tot}} = \int_0^z [\text{Nd}]_z f_{\text{Nd}}^{\text{dust}}(z) \rho(z) dz$$

To obtain an aerosol phosphorus deposition rate, we multiply the integrated Nd flux by the P/Nd ratio of dust sources⁵⁰ and divide by the duration of soil development:

$$P_{\text{dust}} = \text{Nd}_{\text{dust}}^{\text{tot}} \times (P/\text{Nd})_{\text{dust}} \times \Delta t^{-1}$$

This analysis using $^{143}\text{Nd}/^{144}\text{Nd}$, Eu/Eu^* and Th/Hf yields estimates of long-term aerosol phosphorus fluxes ranging from 0.6 to 1.2 $\text{mg P m}^{-2} \text{yr}^{-1}$ to the 150,000-yr-old site (Table 1). We cannot be certain that any given trace element is immobile during weathering, but the similarity among independent estimates of dust flux suggests that we have chosen reliable tracers of dust input. Th and Hf are less mobile than the REE during weathering, yield the highest calculated dust inputs, and may provide the best time-integrated estimate of dust flux. The range of estimates is higher than that calculated using a long-term rate derived from the deep sea record of eolian deposition to the north central Pacific (0.17–0.35 $\text{mg P m}^{-2} \text{yr}^{-1}$)⁵². It is lower than an estimate of modern phosphorus fluxes (2.5 $\text{mg P m}^{-2} \text{yr}^{-1}$) based on direct measurements of mineral aerosol deposition rates in Hawaii and Samoa⁷⁰. The modern numbers have large uncertainties, but they may reflect enhanced atmospheric dust loading due to human activity in Asia.

biological activity in the long run. Second, phosphorus acts as a master regulator of biological activity in the oldest sites on this sequence, in agreement with existing conceptual models for terrestrial biogeochemistry³. It does so not only because the weathering source of phosphorus is depleted, but also because atmospheric inputs of phosphorus are very low in comparison with those of calcium and potassium and other putatively rock-derived elements. And third, even for phosphorus, dust from arid central Asia becomes the dominant source of inputs on a million-year timescale.

We cannot be sure that we have identified all the important sources of element inputs, especially for phosphorus. In particular, locally generated dust from exposed sea cliffs and beaches, and from semi-arid areas within Hawaii might contribute elements to our wet forest sites under some conditions—particularly during the glacial periods when sea level was lower and climates were probably drier³⁵. Also, nesting seabirds could have brought marine phosphorus into

Table 1 Typical endmember values for geochemical tracers

Tracer	Basalt	Dust	Amount of P (mg m^{-2}) delivered in dust
ϵ_{Nd}	+7	−10	0.56
Eu/Eu^*	1.06	0.61	1.04
Th/Hf	0.25	1.84	1.22

The Hawaiian basalt values were measured in the geochemistry laboratory at Cornell (L.A.D. and W. M. White, unpublished results). The estimates of the dust end members are from ref. 57. The final column contains estimates of the flux of P delivered to the 150,000-yr-old site.

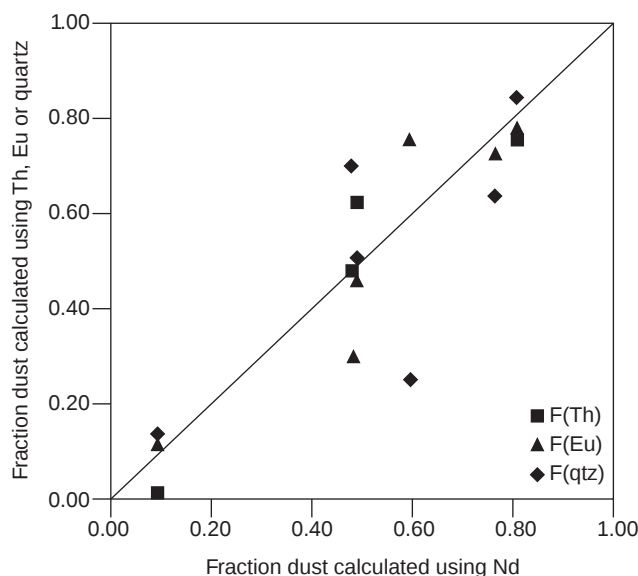


Figure 7 The mass fraction of dust-derived soil component based on Th/Hf ratios, Eu anomalies, and quartz content plotted against the dust mass fraction derived from Nd isotopes for the 150,000-year-old site. General agreement between independent geochemical indicators of dust input suggests that trace element and isotopic ratios can constrain dust inputs even when mineralogical indicators such as quartz content have been lost to dissolution.

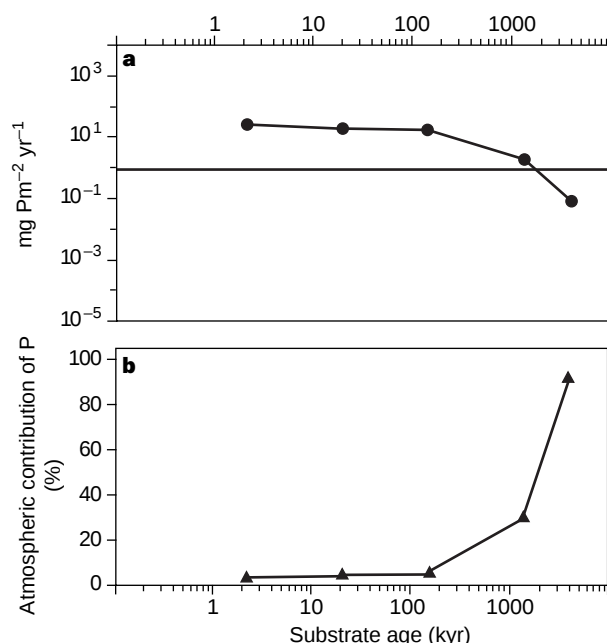


Figure 8 Comparison of phosphorus inputs from substrate and the atmosphere. **a**, Rate of P loss from the soil as a whole (Fig. 3c) plotted with the long-term average addition of P in mineral aerosol. **b**, The per cent atmospheric contribution for each site calculated by dividing the atmospheric contribution by the sum of the atmospheric inputs and contribution from soil phosphorus release.

terrestrial ecosystems, before they were largely extirpated by humans and introduced mammals. These pathways are worth further investigation—but, in any case, it is clear that lava sources decline in importance as a source of elements during the course of ecosystem development, even for phosphorus.

We cannot extrapolate the Hawaiian measurements directly to most continental settings, because both weathering and atmospheric contributions to ecosystems differ in Hawaii compared to most other places. The basaltic substrate is rapidly weathered, making the contribution of weathering to young sites greater in Hawaii than elsewhere, and the contribution to old sites less. More importantly, glaciation resets soil and ecosystem development at relatively frequent intervals over much of the temperate zone, maintaining it at young developmental stages in which weathering is a substantial source of most elements⁶². Tectonically induced erosion can similarly reset soil and ecosystem development anywhere on earth^{63,64}.

For atmospheric inputs, marine aerosol is a less important source of elements in the interior of continents than in Hawaii, and dust is a more important source. For example, Hawaii receives 10–1,000-fold less dust than much of the continental United States⁶⁵, and the rainwater concentrations of Ca, Sr, K and P are larger in continental interiors than over the ocean⁶⁶. On very old, highly weathered substrates in the humid interior of continents (such as the Amazon basin), the combination of less marine aerosol and greater long-distance dust inputs could cause a shift in the limitation of biological processes from phosphorus to the more mobile cations.

Despite these differences, we can use Hawaii to identify some of the general patterns and characterize fundamental mechanisms underlying soil and ecosystem development. Most important, the dependence of biological processes in Hawaii on conditions in, and transport mechanisms from, central Asia demonstrates that the dynamics of long-term soil and ecosystem development cannot be evaluated as a local phenomenon in isolation: nowhere on Earth is that isolated. □

O. A. Chadwick is in the Department of Geography, University of California, Santa Barbara, California 93106, USA; L. A. Derry is in the Department of Geological Sciences, Cornell University, Ithaca, New York 14853, USA; P. M. Vitousek is in the Department of Biological Sciences, Stanford University, Stanford, California 94305, USA; B. J. Huebert is in the Department of Atmospheric Sciences, University of Hawaii, Manoa, Hawaii 96822, USA; and L. O. Hedin is in the Section of Ecology and Systematics, Cornell University, Ithaca, New York 14853, USA.

1. Aber, J. D. & Melillo, J. M. *Terrestrial Ecosystems* (Saunders, Philadelphia, 1991).
2. Schlesinger, W. H. *Biogeochemistry: An Analysis of Global Change* 2nd edn (Academic, San Diego, 1996).
3. Walker, T. W. & Syers, J. K. The fate of phosphorus during pedogenesis. *Geoderma* **15**, 1–19 (1976).
4. Stevens, P. R. & Walker, T. W. The chronosequence concept and soil formation. *Q. Rev. Biol.* **45**, 333–350 (1970).
5. Gorham, E., Vitousek, P. M. & Reiners, W. A. The regulation of element budgets over the course of terrestrial ecosystem succession. *Annu. Rev. Ecol. System.* **10**, 53–84 (1979).
6. Chapin, F. S., Walker, L. R., Fastie, C. L. & Sharman, L. C. Mechanisms of primary succession following deglaciation at Glacier Bay, Alaska. *Ecol. Monogr.* **64**, 149–175 (1994).
7. Cole, C. V. & Heil, R. D. in *Terrestrial Nitrogen Cycles* (eds Clark, F. E. & Ross, T. H.) 363–374 (Ecol. Bul. 33, Stockholm, 1981).
8. Likens, G. E. & Bormann, F. H. *Biogeochemistry of a Forested Ecosystem* 2nd edn (Springer, New York, 1995).
9. Hedin, L. O., Armesto, J. J. & Johnson, A. H. Patterns of nutrient loss from unpolluted, old growth temperate forests: evaluation of biogeochemical theory. *Ecology* **76**, 493–509 (1995).
10. Hedin, L. O. et al. Steep declines in atmospheric base cations in regions of Europe and North America. *Nature* **367**, 351–354 (1994).
11. Tegen, I. & Fung, I. Y. Modeling of mineral dust in the atmosphere: sources, transport, and optical thickness. *J. Geophys. Res.* **99**, 22897–22914 (1994).
12. Swap, R., Garstang, M., Greco, S., Talbot, R. & Kallbert, P. Saharan dust in the Amazon Basin. *Tellus* **44**, 133–149 (1992).
13. Propero, J. M., Glaccum, R. A. & Nees, R. T. Atmospheric transport of soil dust from Africa to South America. *Nature* **289**, 570–572 (1981).
14. Uematsu, M. et al. Transport of mineral aerosol from Asia over the North Pacific Ocean. *J. Geophys. Res.* **88**, 5343–5452 (1983).
15. Lindberg, S. E. & Owens, J. G. Throughfall studies of deposition to forest edges and gaps in montane ecosystems. *Biogeochemistry* **19**, 173–194 (1993).
16. Lovett, G. M. Atmospheric deposition of nutrients and pollutants to North America: an ecological perspective. *Ecol. Appl.* **4**, 629–650 (1994).
17. Weathers, K. C. & Likens, G. E. Clouds in southern Chile: An important source of nitrogen to nitrogen-limited ecosystems. *Environ. Sci. Technol.* **31**, 210–213 (1997).
18. Heath, J. A. & Huebert, B. J. Cloudwater deposition as a source of fixed nitrogen in a Hawaiian montane forest. *Biogeochemistry* (in the press).
19. Cao, M. K. & Woodward, F. I. Dynamic responses of terrestrial ecosystem carbon cycling to global climate change. *Nature* **393**, 249–252 (1998).
20. Aplet, G. H., Hughes, R. F. & Vitousek, P. M. Ecosystem development on Hawaiian lava flows: biomass and species composition. *J. Veg. Sci.* **9**, 17–26 (1998).
21. Austin, A. T. & Vitousek, P. M. Nutrient dynamics on a precipitation gradient in Hawaii. *Oecologia* **113**, 519–529 (1998).
22. Schimel, D. S. et al. Climatic, edaphic, and biotic controls over carbon and turnover of carbon in soils. *Global Biogeochem. Cyc.* **8**, 279–293 (1994).
23. Vitousek, P. M. et al. Human alteration of the global nitrogen cycle: sources and consequences. *Ecol. Appl.* **7**, 737–750 (1997).
24. Aber, J. D. et al. Nitrogen saturation in temperate forest ecosystems: hypothesis revisited. *Bioscience* **48**, 921–934 (1998).
25. Schimel, D. S. et al. Climate and nitrogen control on the geography and timescales of terrestrial biogeochemical cycling. *Global Biogeochem. Cyc.* **10**, 677–692 (1996).
26. Moore, J. G. & Clague, D. A. Volcano growth and evolution of the island of Hawaii. *Geol. Soc. Am. Bull.* **104**, 1471–1484 (1992).
27. Clague, D. A. & Dalrymple, G. B. in *Volcanism in Hawaii* (eds Decker, R. W. et al.) 5–73 (US Geol. Surv. Prof. Paper 1350, 1987).
28. Wright, T. C. & Helz, R. T. in *Volcanism in Hawaii* (eds Decker, R. W. et al.) 625–640 (US Geol. Surv. Prof. Paper 1350, 1987).
29. Macdonald, G. A., Abbott, A. T. & Peterson, F. L. *Volcanoes in the Sea* 2nd edn (Univ. Hawaii Press, Honolulu, 1983).
30. Wagner, W. L., Herbst, D. R. & Sohmer, S. H. *Manual of the Flowering Plants of Hawaii* (Bishop Museum Spec. Publ. 83, Univ. Hawaii Press, Honolulu, 1990).
31. Jenny, H. *Soil Genesis with Ecological Perspectives*. (Springer, New York, 1980).
32. Pickett, S. T. A. in *Long-Term Studies in Ecology: Approaches and Alternatives* (ed. Likens, G. E.) 110–135 (Springer, New York, 1989).
33. Porter, S. C. Hawaiian glacial ages. *Quat. Res.* **12**, 161–187 (1979).
34. Gavenda, R. T. Hawaiian Quaternary paleoenvironments: A review of geological, pedological, and botanical evidence. *Pacific Sci.* **46**, 295–307 (1992).
35. Hotchkiss, S. C. *Quaternary Vegetation and Climate of Hawaii* (thesis, Univ. Minnesota, Minneapolis, 1998).
36. Ludwig, K. R., Szabo, B. J., Moore, J. G. & Simmons, K. R. Crustal subsidence rate off Hawaii determined from ²³⁴U/²³⁸U ages of drowned coral reefs. *Geology* **19**, 171–174 (1991).
37. Torn, M. S., Trumbore, S. E., Chadwick, O. A., Vitousek, P. M. & Hendricks, D. M. Mineral control of soil carbon storage and turnover. *Nature* **389**, 170–173 (1997).
38. Vitousek, P. M. et al. Soil and ecosystem development across the Hawaiian Islands. *GSA Today* **7**(9), 1–8 (1997).
39. Crews, T. et al. Changes in soil phosphorus and ecosystem dynamics across a long soil chronosequence in Hawaii. *Ecology* **76**, 1407–1424 (1995).
40. Vitousek, P. M. & Farrington, H. Nutrient limitation and soil development: experimental test of a biogeochemical theory. *Biogeochemistry* **37**, 63–75 (1997).
41. Kitayama, K. & Mueller-Dombois, D. Vegetation changes during long-term soil development in the Hawaiian montane rainforest zone. *Vegetatio* **120**, 1–20 (1995).
42. Vitousek, P. M., Turner, D. R. & Kitayama, K. Foliar nutrients during long-term soil development in Hawaiian montane rain forest. *Ecology* **76**, 712–720 (1995).
43. Herbert, D. A. *Primary Productivity and Resource use in Metrosideros Polymorpha Forest as Influenced by Nutrient Availability and Hurricane Iniki* (thesis, Univ. Hawaii at Manoa, Honolulu, 1995).

44. Vitousek, P. M., Walker, L. R., Whiteaker, L. D. & Matson, P. A. Nutrient limitation to plant growth during primary succession in Hawaii Volcanoes National Park. *Biogeochemistry* **23**, 197–215 (1993).
45. Herbert, D. A. & Fownes, J. H. Phosphorus limitation of forest leaf area and net primary productivity on a weathered tropical soil. *Biogeochemistry* **29**, 223–235 (1995).
46. Graustein, W. C. in *Stable Isotopes in Ecological research* (eds Rundel, P. W., Ehleinger, J. R. & Nagy, K. A.) 491–512 (Springer, New York, 1989).
47. Capo, R. C., Stewart, B. W. & Chadwick, O. A. Strontium isotopes as tracers of ecosystem processes: Theory and methods. *Geoderma* **83**, 515–524 (1998).
48. Kennedy, M. J., Chadwick, O. A., Vitousek, P. M., Derry, L. A. & Hendricks, D. M. Changing sources of base cations during ecosystem development, Hawaiian Islands. *Geology* **26**, 1015–1018 (1998).
49. Newman, E. I. Phosphorus inputs to terrestrial ecosystems. *J. Ecol.* **83**, 713–726 (1995).
50. Graham, W. F. & Duce, R. A. Atmospheric pathways of the phosphorus cycle. *Geochim. Cosmochim. Acta* **43**, 1195–1208 (1979).
51. Duce, R. A. *et al.* The atmospheric input of trace species of the world ocean. *Global Biogeochem. Cyc.* **5**, 193–259 (1991).
52. Rea, D. K. The paleoclimatic record provided by eolian deposition in the deep sea: the geological history of wind. *Rev. Geophys.* **32**, 159–195 (1994).
53. Parrington, J. R., Zoller, W. H. & Aras, N. K. Asian dust: seasonal transport to the Hawaiian Islands. *Science* **220**, 195–197 (1983).
54. Uematsu, M., Duce, R. A. & Prospero, J. M. Deposition of atmospheric mineral particles to the North Pacific Ocean. *J. Atmos. Chem.* **3**, 123–138 (1985).
55. Jackson, M. L. *et al.* Geomorphological relationships of tropospherically derived quartz in soils of the Hawaiian Islands. *Soil Sci. Soc. Am. J.* **35**, 515–525 (1971).
56. Dymond, J., Biscaye, P. E. & Rex, R. W. Eolian origin of mica in Hawaiian soils. *Geol. Soc. Am. Bull.* **85**, 37–40 (1974).
57. Taylor, S. R. & McLennan, S. M. The geochemical evolution of the continental crust. *Rev. Geophys.* **33**, 241–265 (1995).
58. Nakai, S., Halliday, A. N. & Rea, D. K. Provenance of dust in the Pacific Ocean. *Earth Planet. Sci. Lett.* **119**, 143–157 (1993).
59. Chen, C. Y. & Frey, F. A. Origin of Hawaiian tholeiite and alkalic basalt. *Nature* **302**, 785–789 (1983).
60. Feigenson, M. D. Constraints on the origin of Hawaiian lavas. *J. Geophys. Res.* **91**, 9383–9393 (1986).
61. Ziemann, J. J. *et al.* Atmospheric aerosol trace element chemistry at Mauna Loa Observatory 1. 1979–1985. *J. Geophys. Res.* **100**, 25979–25994 (1995).
62. Birkeland, P. W. *Soils and Geomorphology* (Oxford Univ. Press, New York, 1984).
63. Bull, W. B. *Geomorphic Responses to Climate Change* (Oxford Univ. Press, New York, 1991).
64. Paton, T. R., Humphreys, G. S. & Mitchell, P. B. *Soils: A New Global View* (Yale Univ. Press, New Haven, 1995).
65. Simonson, R. W. Airborne dust and its significance to soils. *Geoderma* **65**, 1–43 (1995).
66. Berner, E. K. & Berner, R. A. *Global Environment: Water, Air, and Geochemical Cycles* (Prentice Hall, Upper Saddle River, NJ, 1996).
67. Brimhall, G. H. *et al.* Deformational mass transport and invasive processes in soil evolution. *Science* **255**, 695–702 (1992).
68. Brimhall, G. H. *et al.* Metal enrichment in bauxites by deposition of chemically mature aeolian dust. *Nature* **333**, 819–824 (1988).
69. DePaolo, D. J. & Wasserburg, G. J. Nd isotopic variations and petrogenic models. *Geophys. Res. Lett.* **3**, 249–252 (1976).
70. Graham, W. F. & Duce, R. A. Atmospheric input of phosphorus to remote tropical islands. *Pacific Sci.* **35**, 241–255 (1981).
71. Chadwick, O. A., Brimhall, G. H. & Hendricks, D. M. From a black to a gray box: a mass balance interpretation of pedogenesis. *Geomorphology* **3**, 369–390 (1990).

Acknowledgements. This research was supported by the Andrew Mellon Foundation, by the NSF, and by NASA-MTPE. For logistical assistance and access to sites, we are indebted to USDA–National Resources Conservation Service, USGS–Biological Resources Division, Hawaii Volcanoes National Park, Hawaii DL&NR, the Nature Conservancy, and Parker Ranch.

Correspondence and requests for materials should be addressed to O.A.C. (e-mail: oac@geog.ucsb.edu).

Scaling and criticality in a stochastic multi-agent model of a financial market

Thomas Lux* & Michele Marchesi†

* Department of Economics, University of Bonn, Adenauerallee 24–42, 53113 Bonn, Germany

† Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy

Financial prices have been found to exhibit some universal characteristics^{1–6} that resemble the scaling laws characterizing physical systems in which large numbers of units interact. This raises the question of whether scaling in finance emerges in a similar way—from the interactions of a large ensemble of market participants. However, such an explanation is in contradiction to the prevalent ‘efficient market hypothesis’⁷ in economics, which assumes that the movements of financial prices are an immediate and unbiased reflection of incoming news about future earning prospects. Within this hypothesis, scaling in price changes would simply reflect similar scaling in the ‘input’ signals that influence them. Here we describe a multi-agent model of financial markets which supports the idea that scaling arises from mutual interactions of participants. Although the ‘news arrival process’ in our model lacks both power-law scaling and any temporal dependence in volatility, we find that it generates such behaviour as a result of interactions between agents.

In our model, the pool of traders is divided into two groups: the first group (‘fundamentalists’) follows the premise of the efficient market hypothesis in that they expect the price (p) to follow the so-called fundamental value of the asset (p_f), which is the discounted sum of expected future earnings (for example, dividend payments). A fundamentalist trading strategy consists of buying (selling) when the actual market price is believed to be below (above) the fundamental value. The second group (called ‘noise traders’ following established terminology in economics⁸), however, does not believe in an immediate tendency of the price to revert to its underlying fundamental value. Instead of focusing on fundamentals, these agents attempt to identify price trends and patterns (charts), and also consider the behaviour of other traders as a source of information, which results in a tendency towards herding behaviour. Furthermore (because it is important for the resulting market operations whether a noise trader believes in a rising or declining market), we further distinguish between optimistic and pessimistic individuals in this group: optimists will buy additional units of the asset, whereas the pessimists will sell part of their actual holdings of the asset.

The main building blocks of the model are movements of individuals from one group to another together with the (exogenous) changes of the fundamental value and the (endogenous) price changes resulting from the agents’ market operations. A distinguishing feature of our approach as compared with other recent simulation models^{9–14} is that we adopt a mass-statistical formalization inspired by statistical physics^{15,16}: individuals react to certain economic forces by changing their behaviour with a certain (endogenous) probability. As a simple formalization of movements into and out of the three groups we use exponential functions, so that a switch from one group to another occurs with a certain endogenous and time-varying probability $ve^{U(t)}\Delta t$ within some small increment Δt . Here the coefficient v is a parameter for the frequency of revaluation of opinion or strategy by the agents (possibly assuming different values for different types of switches), and the function $U(t)$ is a forcing term covering those factors that

are decisive for the pertinent changes of behaviour. The dynamics of our artificial market incorporates the following elements:

Changes of opinion of the noise traders. These changes, from a pessimistic to an optimistic mood and vice versa, are governed by the development of the ‘noise’ factors: the actual price trend is used as a proxy for the influence of ‘chartist’ practices while herding effects are formalized by computing the majority opinion among noise traders as the (scaled) difference between optimistic and pessimistic individuals. A weighted combination of both factors appears in the U function that governs changes of opinion. If both components are in harmony, a dominant trend of switches will ensue. For example, observation of price increases together with a prevalence of optimistic individuals would be seen as a strong indication of a continuation of a rising market, and would result in a tendency of formerly pessimistic individuals to convert to the optimistic group. Conflicting signals, of course, would reduce the disposition to follow either the majority opinion or the actual price trend.

Switches between the noise traders and the fundamentalists. Such switches are driven by the difference between the (momentary) profits earned by individuals in both groups: profits of noise traders from the optimistic group consist in short-term capital gains due to the price change (or losses in the case of a fall of the market price). Fundamentalists, on the other hand, consider the deviation between price and fundamental value as the source of arbitrage opportunities. The difference between both strategies is that the gains of chartists are immediately realized whereas those claimed by fundamentalists occur only in the future and depend on the uncertain time for reversal to the fundamental value. To take account of this, fundamentalists’ arbitrage profits are discounted by a factor <1 . The difference between profits of both groups enters the U -function in the transition probability, with switches from the fundamentalist group to the (optimistic) noise traders dominating if the profit

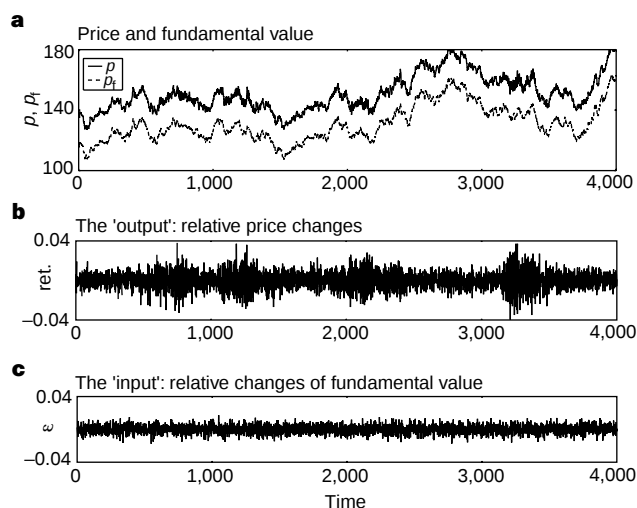


Figure 1 Typical ‘snapshot’ from a longer simulation run. Panel **a** shows the development with time of both the market price p (solid line) and the fundamental value p_f (dotted line). (We have shifted the latter series vertically for better visibility.) Panel **b** shows returns (log changes of the market price: $ret_t = \ln(p_t) - \ln(p_{t-1})$); panel **c** shows log changes of the fundamental value ($\epsilon_t = \ln(p_{f,t}) - \ln(p_{f,t-1})$), respectively; both series have been computed from the time series shown in **a**. The increments of $\ln(p_t)$ follow a normal distribution.

differential is in favour of the latter group and vice versa. For comparison of profits between pessimistic noise traders and fundamentalists, the point of view has to be changed appropriately: because the former rush out of the market in order to avoid losses, their gain is given by the difference between the average profit rate from alternative investments (assumed to be constant) minus the price change (which, when negative, amounts to a capital loss) of the asset they sell. Again, a profit differential in favour of one of the two groups tends to induce changes of behaviour among members of the other group.

Price changes. These are endogenous responses of the market to imbalances between demand and supply—excess demand (supply) leading to an increase (decrease) of the prevailing price. Demand and supply, however, originate from the decisions of our agents: assuming a constant average trading volume of noise traders, their demand and supply is readily determined by the actual numbers of optimistic and pessimistic individuals. Fundamentalists' sensitivity to the relative deviation of the price from the fundamental value, on the other hand, amounts to an excess demand of this group depending on the difference $p - p_f$. Overall excess demand is the sum of both components.

Changes of fundamental value. These constitute the external driving force which affects the market through the operations of fundamentalist traders. In order to ensure that none of the typical characteristics of financial prices can be traced back to exogenous factors, we assume that the relative changes of p_f are Gaussian random variables, i.e. $\ln(p_{f,t}) - \ln(p_{f,t-1}) = \epsilon_t$ with ϵ_t following a normal distribution with mean zero and time-invariant variance σ_ϵ^2 .

A more detailed description of our model and some theoretical results are available as Supplementary Information. Theoretical analysis reveals that stationary states of our dynamics are characterized by a price which on average equals the fundamental value. Hence, at least in the long term, we have an 'efficient' market which

incorporates all new information into market prices: no 'pathological' situations, with persistent deviations from the economy's fundamentals, occur. A typical simulation (Fig. 1a) shows how closely the price tracks the development of the fundamental value. But comparing the time paths of returns extracted from the price path: $\text{ret.}(\tau) = \ln(p_t) - \ln(p_{t-\tau})$ with relative changes of p_f (Fig. 1b and c) it is evident that the distributional characteristics of ϵ_t are not reflected in similarly normally distributed returns—despite the close association of the integrated time paths in Fig. 1a, the statistical properties of the increments differ fundamentally. In particular, the time series of returns exhibits a higher frequency of extreme events and clustering of volatility.

Figure 2 shows the differences of the unconditional distributions between the input (logarithm of changes of p_f) and the output (relative price changes). As compared with the exponential fall-off of the density of the input, one observes a clear widening of the distribution of large price fluctuations which roughly follows a power law. Determining the exponent α of the Pareto distribution for the tails (that is, $F(\text{ret.} > x) \approx cx^{-\alpha}$) from a regression in logarithmic coordinates yields an estimate of 2.64 for data with a unit time step; this is in good agreement with results obtained for empirical data at daily frequencies^{17,18}. However, for returns at lower frequencies (that is, under time aggregation), we also observe a cross-over to the normal distribution with increasing time lag τ . The same happens for empirical financial data.

Turning to the issue of temporal dependence: we estimated the self-similarity parameter H for both raw returns and absolute returns using the approach of Peng *et al.*¹⁹ (Fig. 3). For raw returns, differences in scaling behaviour between the input and output time series are small, yielding $H = 0.49$ and $H = 0.48$, respectively. This is in accordance with absence of long memory ($H = 0.5$) in empirical financial returns. As a consequence, the degree of predictability of price changes is small. However, the picture changes

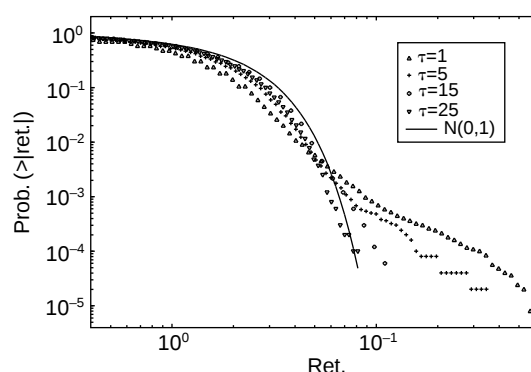


Figure 2 Log-log plot of the complement of the cumulative distribution of returns ($\text{ret.}(\tau) = \ln(p_t) - \ln(p_{t-\tau})$) at different levels of time aggregation: $\text{ret.}(\tau) = \ln(p_t) - \ln(p_{t-\tau})$. All time series are scaled by their sample standard deviation and the positive and negative tails have been merged by using absolute returns. For comparison, the solid line gives the complement of the cumulative distribution of the standard normal distribution on which the scaled changes of the fundamental value, $\epsilon_t = \ln(p_{f,t}) - \ln(p_{f,t-1})$, would collapse at all levels of aggregation. For the highest frequencies, one observes clear deviations from the exponential decay with approximate power-law scaling. Performing a log-log regression on the largest 30% of the observations, the estimated slope is -2.64 ± 0.077 at unit time steps ($\tau = 1$). This is close to results obtained for various financial prices at daily frequencies. Increasing the time step τ , a cross-over to the normal distribution is observed. This is also known to occur with financial data when proceeding from daily frequency to lower (weekly, monthly) frequencies.

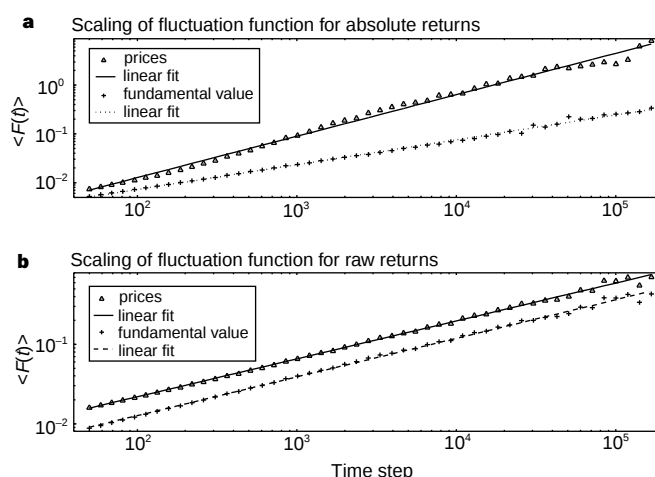


Figure 3 Estimation of self-similarity parameter H . **a**, Absolute returns; **b**, raw returns. Using the approach of Peng *et al.*¹⁹ the exponent H was estimated from the behaviour of the average fluctuation $\langle F(t) \rangle$ of a random variable over its local trend in intervals of size t . The expected behaviour is a power law, $\langle F(t) \rangle \propto t^H$, from which H can be extracted performing a regression in log coordinates. Panel **b** compares the scaling of raw returns and changes of $\ln(p_f)$. For the latter, the self-similarity parameter H is estimated to be $H = 0.49 \pm 0.03$ (slope of dashed line), which is close to the theoretically expected value of 0.5 for a white-noise process. The scaling of returns yields $H = 0.48 \pm 0.003$, which differs only slightly from the results for the input. (The curve for p_f has been shifted vertically for better visibility.) Panel **a** depicts the development of the fluctuation function of absolute returns. Here we see clear differences between the behaviour of the exogenous force ($H = 0.51 \pm 0.004$, the slope of the dotted line) and the series of absolute returns, the latter being characterized by strong persistence with estimated $H = 0.85 \pm 0.010$ (solid line).

dramatically when considering absolute returns as a measure of volatility. Here we see that the transformed price data behave differently from their counterpart derived from the input series, and exhibit $H = 0.85$ which is a sign of strong persistence in volatility. The exponent is again very close to the scaling found for empirical data^{6,20}.

As these scaling properties are absent in the external driving force, they are generated by the interaction of economic agents with heterogeneous beliefs and strategies in our simulated market. Can we explain the emergence of power laws in these simulations? A closer investigation reveals that the alternation between tranquil and turbulent periods comes about through the changes of agents between groups. In particular, in periods of high volatility we also find a large fraction of agents in the noise trader group. Theoretical analysis shows that a critical value for the number of noise traders exists where the system loses its stability. Volatility is above average when the fraction of noise traders comes close to this critical point, or even increases beyond it, for some time. However, the ensuing destabilization is only temporary, and turbulent phases are overcome quickly by endogenous mechanisms: large deviations from the fundamental value are seen as profit opportunities by fundamentalists whose operations then tend to stabilize the market again. This temporal instability is similar to mechanisms found recently in various models in physics where this phenomenon has been denoted on-off intermittency^{21–23}. As the possibility of temporal destabilization (with ensuing bursts of volatility) exists for an open set of parameter values, the qualitative outcome of our model seems to be extremely robust. This has been confirmed by simulations with many different parameter sets (data not shown), which all led to endogenous emergence of power-law tails and temporal dependence of volatility, albeit with varying coefficients α and H . $\square$

Received 29 October; accepted 27 November 1998.

- Mandelbrot, B. The variation of certain speculative prices. *J. Bus.* **35**, 394–419 (1963).
- Fama, E. Mandelbrot and the stable Paretian hypothesis. *J. Bus.* **35**, 420–429 (1963).
- Mantegna, R. N. & Stanley, E. Scaling behaviour in the dynamics of an economic index. *Nature* **376**, 46–49 (1995).
- Ghashgaie, S., Breymann, W., Peinke, J., Talkner, P. & Dodge, Y. Turbulent cascades in foreign exchange markets. *Nature* **381**, 767–770 (1996).
- Galluccio, S., Caldarelli, G., Marsili, M. & Zhang, Y.-C. Scaling in currency exchange. *Physica A* **245**, 423–436 (1997).
- Liu, Y., Cizeau, P., Meyer, M., Peng, C.-K. & Stanley, H. E. Correlations in economic time series. *Physica A* **245**, 437–440 (1997).
- Fama, E. Efficient capital markets: a review of theory and empirical work. *J. Fin.* **25**, 383–417 (1970).
- Shleifer, A. & Summers, L. H. The noise trader approach to finance. *J. Econ. Perspect.* **4**, 19–33 (1990).
- Palmer, R. G., Arthur, W. B., Holland, J. H., LeBaron, B. & Taylor, P. Artificial economic life: a simple model for a stockmarket. *Physica D* **75**, 264–274 (1994).
- Levy, M., Levy, H. & Solomon, S. Microscopic simulation of the stock market: the effect of microscopic diversity. *J. Phys. I (France)* **5**, 1087–1107 (1995).
- Bak, P., Paczuski, M. & Shubik, M. Price variations in a stock market with many agents. *Physica A* **246**, 430–453 (1997).
- Caldarelli, G., Marsili, M. & Zhang, Y.-C. A prototype model of stock exchange. *Europhys. Lett.* **40**, 479–484 (1997).
- Brock, W. & LeBaron, B. A dynamical structural model for stock return volatility and trading volume. *Rev. Econ. Stat.* **78**, 94–110 (1996).
- Brock, W. & Hommes, C. Rational routes to randomness. *Econometrica* **65**, 1059–1095 (1997).
- Lux, T. Time variation of second moments from a noise trader/infection model. *J. Econ. Dyn. Control* **22**, 1–38 (1997).
- Lux, T. The socio-economic dynamics of speculative markets: interacting agents, chaos, and the fat tails of return distributions. *J. Econ. Behav. Organizat.* **33**, 143–165 (1998).
- Guillaume, D. M. *et al.* From the bird's eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange markets. *Fin. Stoch.* **1**, 95–129 (1997).
- Gopikrishnan, P., Meyer, M., Amaral, L. A. N. & Stanley, H. E. Inverse cubic law for the distribution of stock price variations. *Eur. Phys. J. B* **3**, 139–140 (1998).
- Peng, C.-K. *et al.* Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49**, 1685–1689 (1994).
- Lux, T. Long-term stochastic dependence in financial prices: evidence from the German stock market. *Appl. Econ. Lett.* **3**, 701–706 (1996).
- Platt, N., Spiegel, E. A. & Tresser, C. On-off intermittency: a mechanism for bursting. *Phys. Rev. Lett.* **70**, 279–282 (1993).
- Haeghe, J. F., Platt, N. & Hammel, S. M. Characterization of on-off intermittency. *Phys. Rev. E* **49**, 1140–1150 (1994).
- Youssefmir, M. & Huberman, B. Clustered volatility in multiagent dynamics. *J. Econ. Behav. Organizat.* **32**, 101–118 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. Financial support by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303 at the University of Bonn is acknowledged.

Correspondence and requests for materials should be addressed to T.L. (e-mail: lux@iwi.uni-bonn.de).

A single-photon turnstile device

J. Kim*, O. Benson*, H. Kan† & Y. Yamamoto*‡

* ERATO Quantum Fluctuation Project, Edward L. Ginzton Laboratory, Stanford University, Stanford, California 94305, USA

† ERATO Quantum Fluctuation Project, Hamamatsu Photonics Inc., Hamamatsu, Shizuoka, 434-0041, Japan

‡ NTT Basic Research Laboratories, 3-1 Morinosato-Wakamiya Atsugi, Kanagawa, 243-01, Japan

Quantum-mechanical interference between indistinguishable quantum particles profoundly affects their arrival time and counting statistics. Photons from a thermal source tend to arrive together (bunching) and their counting distribution is broader than the classical Poisson limit¹. Electrons from a thermal source, on the other hand, tend to arrive separately (anti-bunching) and their counting distribution is narrower than the classical Poisson limit^{2–4}. Manipulation of quantum-statistical properties of photons with various non-classical sources is at the heart of quantum optics: features normally characteristic of fermions—such as anti-bunching, sub-Poissonian and squeezing (sub-shot-noise) behaviours—have now been demonstrated⁵. A single-photon turnstile device was proposed^{6–8} to realize an effect similar to conductance quantization. Only one electron can occupy a single state owing to the Pauli exclusion principle and, for an electron waveguide that supports only one propagating transverse mode, this leads to the quantization of electrical conductance: the conductance of each propagating mode is then given by $G_Q = e^2/h$ (where e is the charge of the electron and h is Planck's constant; ref. 9). Here we report experimental progress towards generation of a similar flow of single photons with a well regulated time interval.

When a light-emitting p–n junction is driven with a high-impedance constant-current source, injection of electron–hole pairs can be regulated to below the classical shot-noise limit and light with sub-shot-noise intensity fluctuations can be generated¹⁰. This is possible because the inelastic scattering of electrons in a highly dissipative resistor can suppress the current noise by means of the Pauli exclusion principle^{11,12}, and the Coulomb repulsive interaction between electrons in a p–n junction can suppress the electron injection noise by way of the collective Coulomb blockade effect^{13–15}. In these squeezing experiments with a macroscopic p–n junction, however, only large numbers of photons (of the order of $\sim 10^8$) can be regulated owing to a small single charging energy.

It has been demonstrated in mesoscopic physics that an ultra-small tunnel junction regulates the electron transport one by one owing to a single charging energy that is large compared to the thermal background energy^{16–18}. If such a single-electron control technique could be extended to simultaneous control of electron and hole in a p–n junction, a single photon would be regularly emitted, one by one⁶.

Our single-photon turnstile device utilizes simultaneous Coulomb blockade for electrons and holes in a mesoscopic double barrier p–n junction (Fig. 1a). The structure consists of an intrinsic central quantum well (QW) in the middle of a p–n junction and the n-type and p-type side QWs isolated by tunnel barriers from the central QW. The lateral size of the device is reduced to increase the single charging energy $e^2/2C_i$, where C_i (i is n or p) is the capacitance between the central QW and the i -side QW. At a certain bias voltage V_0 , the conditions for electron resonant tunnelling are fulfilled, and the m th electron can tunnel into an electron sub-band in the central QW. When the m th electron tunnels, the Coulomb repulsive interaction between electrons shifts

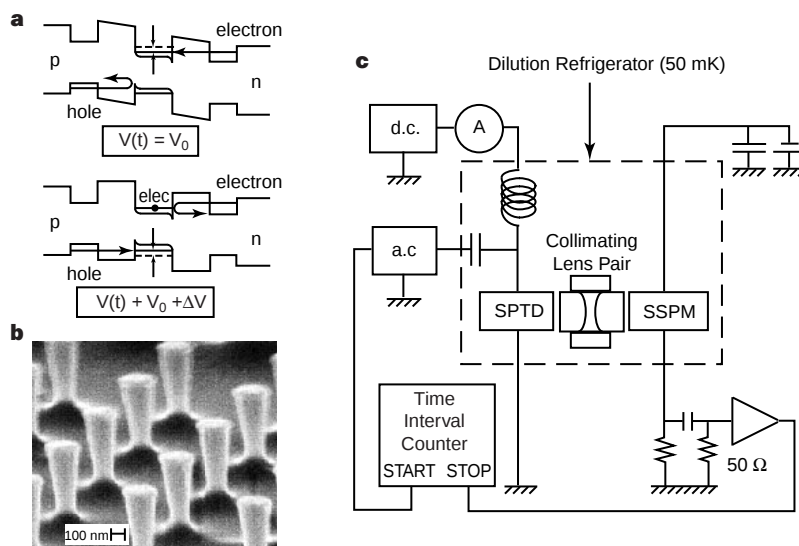


Figure 1 Operation and fabrication of our single-photon turnstile device. **a**, The operational principle of our device. When a junction voltage satisfies the electron resonant-tunnelling condition ($V = V_0$), a single (m th) electron tunnels into the central QW, and the increased Coulomb repulsive interaction between electrons prohibits subsequent electron tunnelling. Throughout the process, the hole tunnelling remains off resonant. When the junction voltage is switched to the hole resonant-tunnelling condition ($V = V_0 + \Delta V$), a single (first) hole is allowed to tunnel, but further hole tunnelling is prohibited owing to the decreased Coulomb attractive interaction between electrons and holes. Throughout this process, the electron tunnelling remains off resonant. The junction voltage is periodically modulated between these two resonant conditions so that a single electron and a single hole are injected per modulation period to produce a single photon. **b**, Scanning electron micrograph of typical etched post structures. The diameters

of the devices range between 200 and 1,000 nm, and the height of the posts is ~ 700 nm. The junctions are located 600 nm from the top, near the bottom of the post structure. In the experiment, electrical connection is made to only one of these posts. **c**, Schematic of the experimental set-up. The fabricated single-photon turnstile device (SPTD) was installed in the mixing chamber of a dilution refrigerator at ~ 50 mK. The device was biased with both a d.c. and an a.c. voltage source. Photons generated from the device were focused through a collimating lens pair onto the surface of a single photon counting detector based on a solid state photomultiplier (SSPM). The temperature of the SSPM was maintained at 6.5 K with good thermal isolation. The time delay between the rising edge of the modulation input and the photon detection event was measured using a time interval counter.

the electron sub-band energy to above the Fermi level of the n-side QW, so the $(m + 1)$ -th electron tunnelling is inhibited. In our single-photon turnstile device, the number of electrons in the central quantum well is $m \approx 10$ at an operating bias condition. At this bias voltage, the hole resonant-tunnelling condition is not satisfied (the Fermi level of the p-side QW is higher than the hole sub-band energy of the central QW), so there is no hole in the central QW. Then the bias voltage is increased to $V_0 + \Delta V$ to satisfy

the hole resonant-tunnelling condition. If a single hole tunnels into the hole sub-band of the central QW, the negative charge of this well is decreased by one and the subsequent hole tunnelling is inhibited owing to the decreased Coulomb attractive interaction between electrons and holes. By periodically modulating the bias voltage between V_0 and $V_0 + \Delta V$, we can inject periodically a single (m th) electron and a single (first) hole into the central QW (Fig. 1 of Supplementary Information). If the tunnel time and the radiative

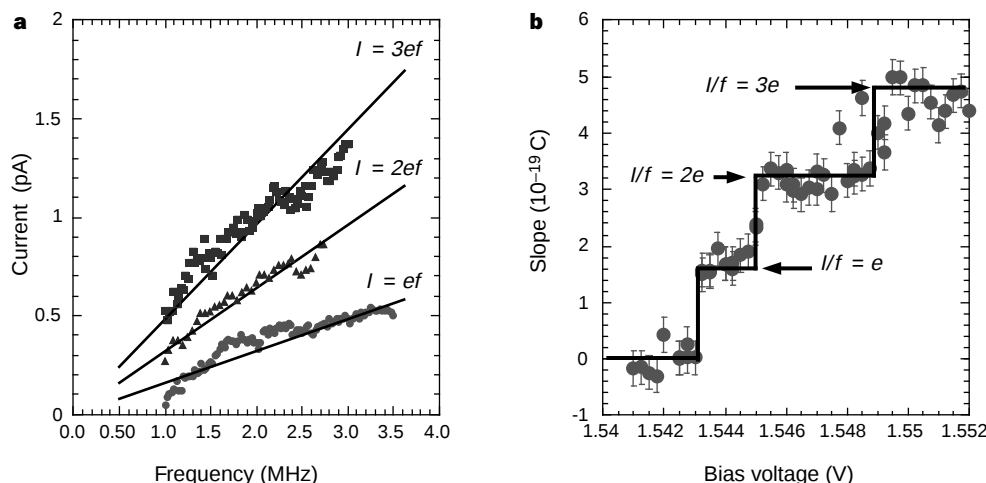


Figure 2 The electrical characteristics of the single-photon turnstile device. **a**, The modulation-frequency dependence of the direct current in the 600-nm turnstile device. A fixed a.c. amplitude of 72 mV is superposed on the three different d.c. voltages, $V = 1.545$, 1.547 and 1.550 V. The background direct (leakage) current, which is independent of the modulation frequency, has been subtracted. The

measured current agrees with the relation $I = nef$ (solid lines), where e is the electron charge and $n = 1, 2$ and 3. **b**, The slopes I/f in the current-frequency curve versus d.c. bias voltage. The slopes I/f are quantized and form plateaux at multiples of e , indicating the operations at $I = ef$, $2ef$ and $3ef$.

recombination time of an electron–hole pair are much shorter than the pulse duration, one (and only one) photon is emitted for every modulation period.

A GaAs/AlGaAs three-QW structure sandwiched by n-type and p-type AlGaAs bulk layers was grown using a molecular beam epitaxy technique. Post structures with diameters of 200–1,000 nm were made by electron-beam lithography followed by metal evaporation, lift-off, and BCl_3/Cl_2 electron cyclotron resonance plasma etching. A scanning electron micrograph of typical etched posts is shown in Fig. 1b. The surface of the device was passivated with sulphur in ammonium sulphide $((\text{NH}_4)_2\text{S})$ solution, and encapsulated by silicon nitride film¹⁹. The structure was covered with hard-baked photoresist and the top of the post was exposed by O_2 plasma etching. Finally, electrical connection was made to each one of the posts independently. The top semi-transparent metal served as the p-type contact from which an emitted photon is detected, and the n-type contact was formed in the substrate.

The device was installed in a dilution refrigerator with base temperature of ~ 50 mK, and was biased with a d.c. and an a.c. voltage source (Fig. 1c). A direct current flowing through the device was measured as a function of d.c. bias voltage with a square wave a.c. modulation voltage. The emitted photon from this device was detected by a Si solid-state photomultiplier (SSPM). This detector features a high quantum efficiency of $\sim 88\%$, a high multiplication gain of $\sim 30,000$, a fast response time of ~ 2 ns and no multiplication noise²⁰. The detector was installed on the mixing chamber of the dilution refrigerator, but the temperature was held at 6.5 K with good thermal isolation.

We measured the d.c. current–voltage characteristics and observed a well-defined resonant-tunnelling current peak with very small background current; this indicates that surface (leakage) current is well suppressed, in spite of the very small size of the post,

by the above-mentioned passivation process (Fig. 2 of Supplementary Information). Measurement of the photons generated by this device indicated that the internal efficiency of photon emission from an electron–hole pair is $\geq 33\%$.

When an a.c. modulation voltage was applied to the device in addition to the d.c. bias voltage, we observed that the direct current increased linearly as a function of the modulation frequency, f . Figure 2a shows the measured current as a function of a.c. modulation frequency, with a fixed a.c. amplitude of 72 mV at three different d.c. bias voltages for a device with a diameter of 600 nm. The measured current (I) was in close agreement with the relation $I = ef$, $I = 2ef$ and $I = 3ef$ (solid lines), when a frequency-independent background current was subtracted. This background current varies from device to device, and ranges from 0.5 to 6.5 pA. In Fig. 2b we evaluate the slope I/f from the current versus frequency curves and plot it as a function of the d.c. bias voltage. We find that the slope increases discretely, creating plateaux at $I/f = ne$, where $e = 1.6 \times 10^{-19}$ C is the charge of an electron and $n = 1, 2$ and 3.

The locking of the current at multiples of the modulation frequency ($I = nf$) suggests that the charge transfer through the device is strongly correlated with the external modulation signal^{16–18}. At the first current plateau at $I = ef$, a single (m th) electron and a single (first) hole are injected into the central QW per modulation period, resulting in single-photon emission. At the second current plateau at $I = 2ef$, two (m th and $(m+1)$ -th) electrons and two (first and second) holes are injected into the central QW per modulation period, resulting in up to two-photon emission (Fig. 3 of Supplementary Information). Similarly, at the third current plateau at $I = 3ef$, three electrons and three holes are injected per modulation period, resulting in up to three-photon emission. This multiple charge operation becomes possible because of the relatively broad inhomogeneous linewidths of the n-side and

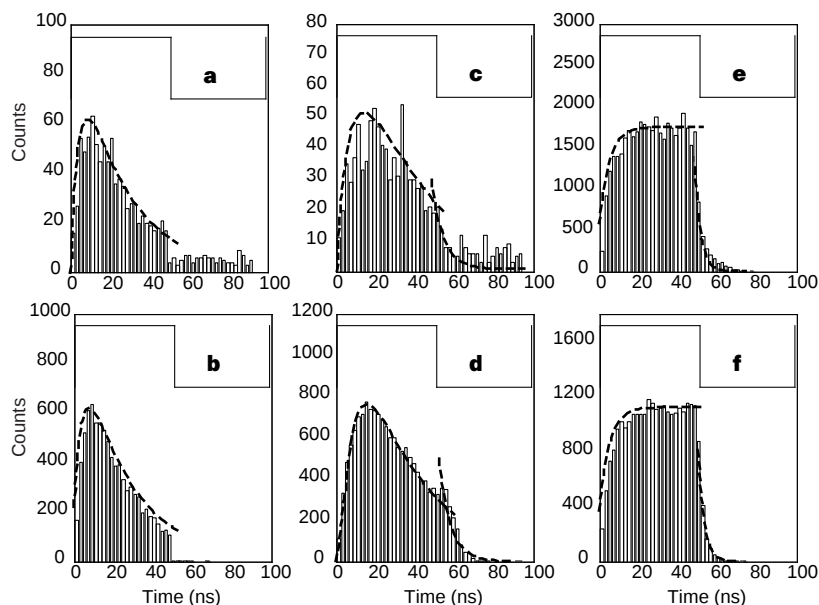


Figure 3 The photon emission characteristics of the single-photon turnstile device. **a**, Histogram of the measured time delay between the rising edge of the modulation input and the photon detection event at the first plateau ($I = ef$). The solid line at the top indicates the external a.c. modulation voltage. As a single electron is injected at V_0 and a single hole is injected at $V_0 + \Delta V$, a photon is emitted after the rising edge of the pulse. The dashed line is an analytical solution (see text). The finite photon counts during the off-pulse period are due to the dark counts of the Si SSPM. **b**, Results of a Monte-Carlo numerical simulation for the photon emission probability versus the time delay at the first plateau ($I = ef$). **c**, Measured histogram of the time delay at the second plateau ($I = 2ef$). The distribution is broader, as two electrons and two holes are injected, and up to

two photons are generated per modulation period. The two dashed lines are the theoretical solutions (see text). The sharp cut-off of photon emission after the falling edge of the modulation input is due to the decay of the hole population due to simultaneous electron–hole recombination and reverse hole tunnelling. **d**, As **b**, but for the second plateau. **e**, Measured histogram of the time delays for a larger-area device (diameter, $1.4 \mu\text{m}$) at a higher temperature (4 K), where the Coulomb blockade effect is absent. The recombination time is shorter in this case (~ 5 ns) due to higher carrier density at the operating condition. **f**, Monte-Carlo numerical simulation result for this case of a modulated classical light-emitting diode.

p-side QWs. The experimental results were well reproduced by Monte-Carlo numerical simulations with a finite resonance line-width (data not shown).

To observe the temporal correlation between the modulation input and photon emission, we measured the time delay from the rising edge of the modulation input to the photon detection event at the first current plateau ($I = ef$) and at the second current plateau ($I = 2ef$). The probability of a single electron-hole pair injected to the central QW of the turnstile device being detected as a photon in the detector was $\sim 1 \times 10^{-4}$ owing to a poor optical coupling efficiency between the two devices in the present set-up. However, the detection quantum efficiency does not affect the time correlation characteristics. Histograms of the measured time delay with 10-MHz modulation frequency are shown in Fig. 3a (for $I = ef$) and Fig. 3c (for $I = 2ef$). The photon emission probabilities have peaks near the rising edge of the modulation input. The rapid increase of the photon emission probability is associated with the hole tunnelling time ($\tau_h \approx 4$ ns), and the slow decay of the photon emission probability corresponds to the electron-hole recombination lifetime ($\tau_{ph} \approx 25$ ns: due to both radiative and non-radiative processes). The photon emission probability in Fig. 3a decays to a non-zero value during the on-pulse due to photons generated by background current. The ratio of the counts contained in the peak to those contained in the non-zero background is $\sim 3:1$, consistent with the ratio of the turnstile current to the background current in this device. The second and faster decay for $I = 2ef$ (Fig. 3c) represents the decay of the hole via backward tunnelling and electron-hole recombination. The associated lifetime for this decay is $(\tau_h^{-1} + \tau_{ph}^{-1})^{-1}$. The dashed lines show the analytical solutions using these parameters.

The experimental results, as well as the analytical traces, are well reproduced by Monte Carlo numerical simulations, as shown in Fig. 3b ($I = ef$) and Fig. 3d ($I = 2ef$). The fact that the photon emission probability decreases during the duration of the on-pulse is a unique indication that the number of holes injected during an on-pulse is restricted to either one or two due to the Coulomb blockade effect (Fig. 4 of Supplementary Information). Figure 3e shows the experimental result for a larger-area device at a higher temperature of 4 K, where the Coulomb blockade effect is absent. In this case, arbitrary numbers of holes are allowed to tunnel into the central QW during an on-pulse, and so the resulting photon emission probability should increase monotonically with a time constant τ_{ph} to the steady-state value. This result is well-reproduced by the simulation (Fig. 3f).

The experimental results reported here provide evidence for the long-sought generation of regulated streams of single photons (and n -photons) where the time interval between single photons is regulated to beyond the Poisson limit. If the un-regulated background photons were suppressed and the collection efficiency was improved, such a non-classical photon source could provide an efficient source for future quantum information technologies, and could also be a useful tool for fundamental tests of quantum mechanics. $\square$

Received 5 August; accepted 7 December 1998.

1. Hanbury Brown, R. & Twiss, R. Q. Correlation between photons in two coherent beams of light. *Nature* **177**, 27–29 (1956).
2. Reznikov, M., Heiblum, M., Shtrikman, H. & Mahalu, D. Temporal correlation of electrons: suppression of shot noise in a ballistic quantum point contact. *Phys. Rev. Lett.* **75**, 3340–3343 (1995).
3. Kumar, A., Saminadayar, L., Glattli, D. C., Jin, Y. & Etienne, B. Experimental test of the quantum shot noise reduction theory. *Phys. Rev. Lett.* **76**, 2778–2781 (1996).
4. Liu, R. C., Odom, B., Yamamoto, Y. & Tarucha, S. Quantum interference in electron collision. *Nature* **391**, 263–265 (1998).
5. Walls, D. F. & Milburn, G. J. *Quantum Optics* (Springer, Berlin, 1994).
6. Imamoglu, A. & Yamamoto, Y. Turnstile device for heralded single photons: Coulomb blockade of electron and hole tunneling in quantum confined p - i - n heterojunctions. *Phys. Rev. Lett.* **72**, 210–213 (1994).
7. Imamoglu, A., Schmidt, H., Woods, G. & Deutsch, M. Strongly interacting photons in a nonlinear cavity. *Phys. Rev. Lett.* **79**, 1467–1470 (1997).
8. Yamamoto, Y. A photon in solitary confinement. *Nature* **390**, 17–18 (1997).
9. von Klitzing, K., Dorda, G. & Pepper, M. New method for high-accuracy determination of the fine-structure constant based on quantized Hall resistance. *Phys. Rev. Lett.* **45**, 494–497 (1980).

10. Yamamoto, Y. in *Quantum Optics of Confined Systems* (eds Ducloy, M. & Bloch, D.) 201–281 (Kluwer, Dordrecht, 1996).
11. Beenakker, C. W. J. & Büttiker, M. Suppression of shot noise in metallic diffusive conductors. *Phys. Rev. B* **46**, 1889–1892 (1992).
12. Liu, R. C. & Yamamoto, Y. Suppression of quantum partition noise in mesoscopic electron branching circuits. *Phys. Rev. B* **49**, 10520–10532 (1994).
13. Imamoglu, A. & Yamamoto, Y. Noise suppression in semiconductor p - i - n junctions: Transition from macroscopic squeezing to mesoscopic Coulomb blockade of electron emission process. *Phys. Rev. Lett.* **70**, 3327–3330 (1993).
14. Kim, J., Kan, H. & Yamamoto, Y. Macroscopic Coulomb-blockade effect in a constant-current-driven light-emitting diode. *Phys. Rev. B* **52**, 2008–2012 (1995).
15. Kim, J. & Yamamoto, Y. Theory of noise in p - n junction light emitters. *Phys. Rev. B* **55**, 9949–9959 (1997).
16. Delsing, P., Likharev, K. K., Kuzmin, L. S. & Claeson, T. Time-correlated single-electron tunneling in one-dimensional arrays of ultrasmall tunnel junctions. *Phys. Rev. Lett.* **63**, 1861–1864 (1989).
17. Geerligs, L. J. et al. Frequency-locked turnstile device for single electrons. *Phys. Rev. Lett.* **64**, 2691–2694 (1990).
18. Kouwenhoven, L. P. et al. Quantized current in a quantum-dot turnstile using oscillating tunnel barriers. *Phys. Rev. Lett.* **67**, 1626–1629 (1991).
19. Hobson, W. S. et al. Silicon nitride encapsulation of sulfide passivated GaAs/AlGaAs microdisk lasers. *J. Vac. Sci. Technol. A* **13**, 642–645 (1995).
20. Kim, J., Yamamoto, Y. & Hogue, H. H. Noise-free avalanche multiplication in Si solid state photomultipliers. *Appl. Phys. Lett.* **70**, 2852–2854 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank H. H. Hogue for providing us with SSPM detectors. This work was partially supported by JSEP.

Correspondence and requests for materials should be addressed to Y.Y. (yamamoto@loki.stanford.edu).

Modulated phases and proton centring in ice observed by X-ray diffraction up to 170 GPa

Paul Loubeyre^{*†}, René LeToullec[†], Elodie Wolanin[†], Michel Hanfland[‡] & Daniel Hausermann[‡]

^{*}Laboratoire Etats Extrêmes Statiques, SPMC/DPTA, CEA 91680 Bruyères-le-Châtel, France

[†]CNRS and Université Paris 6, PMC boîte 77, 4 place Jussieu 75252 Paris, France

[‡]ESRF, BP 220, 38043 Grenoble, France

Because of its open hydrogen-bonded structure, ice shows many structural changes between different crystalline forms under high pressure. Crystallographic studies of these transitions have been pursued largely by neutron scattering, which allows the positions of the hydrogen atoms to be identified^{1,2}. Such studies have previously been extended to pressures of up to 20 GPa, which is however insufficient to permit the investigation of ice X, a 'symmetric ice' in which the protons are thought to reside midway between the oxygen atoms^{3–5}. So far, information about ice X has therefore come from indirect methods such as infrared^{6,7} or Brillouin⁸ spectroscopy. Here we show that single-crystal X-ray diffraction is able to reveal the signature of hydrogen-bond symmetrization. The 111 reflection can be assigned to the hydrogen atoms alone, and we can measure it up to 170 GPa in a diamond anvil cell. This diffraction line (normalized against the intensity of the 222 line, which is due mostly to oxygen atoms) indicates that the proton centring in ice X occurs from about 60 to 150 GPa; at this latter pressure the intensity increases sharply, signalling a further structural change. At lower pressures, we see ice VII ordering in a sequence of spatially modulated phases between 2.2 and 25 GPa, which suggests an analogy with the incommensurate phases of the frustrated Ising model⁹.

We use the single-crystal X-ray diffraction technique with the high brilliance of the European Synchrotron Radiation Facility to measure the structural properties of ice VII in a diamond anvil cell at very high pressure, as described previously for solid hydrogen¹⁰. Ice VII is the stable form of ice at ambient temperature above 2.2 GPa, but its detailed nature remains unknown, with possible multisite disorder of both the oxygen and hydrogen atoms². The single crystal was grown by looking optically at the solid–fluid

equilibrium on the melting curve of ice VII, with no preferential orientation for the growth. Embedding the single crystal in a quasihydrostatic cushion—here we used gold in a ring around the crystal rather than He because of the formation of clathrates—is necessary to significantly reduce its fragmentation up to the 100 GPa range. Good hydrostatic conditions could be achieved up to 20 GPa, but non-hydrostatic compression was observed at higher pressures. Three experiments were performed to determine the equation of state (EOS) of ice at very high pressure. At least six classes of reflections were observed up to 170 GPa and related by the cubic orientation matrix as $\langle 110 \rangle$, $\langle 220 \rangle$, $\langle 111 \rangle$, $\langle 222 \rangle$, $\langle 200 \rangle$ and $\langle 121 \rangle$. The diffraction data are consistent with the body-centred-cubic oxygen sublattice persisting up to 170 GPa. In Fig. 1, the single-crystal measurements of the lattice parameter of the cubic unit cell are compared with those performed on a powdered sample up to a maximum pressure of 128 GPa (refs 11, 12). The good agreement between these different sets of data gives confidence that a correct determination of the EOS of ice can be reached, even though

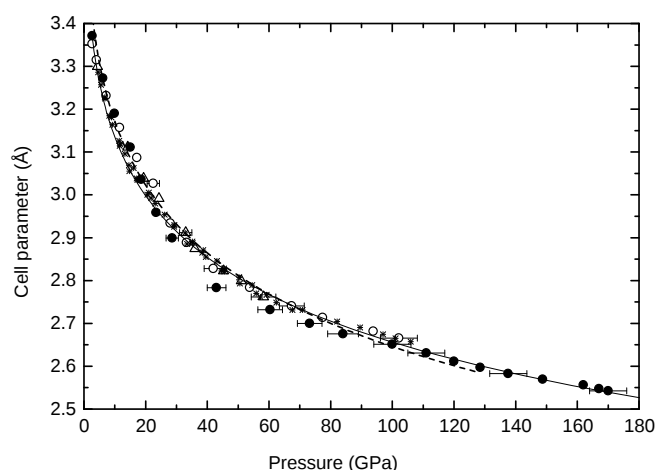


Figure 1 The pressure-cell parameter relation in ice at 300 K. Filled circles, open circles and open triangles indicate results of three single-crystal experiments; asterisks show measurements on powdered ice¹². The dashed line is a good fit to data from ref. 11 up to 128 GPa. A Vinet form¹⁷—promoted as a universal equation of state for solids under strong compression—is satisfactorily applied here to represent the data within the dispersion of the various determinations, shown as the solid line ($V = a^3/2$). It is given by:

$$P = 3K_0 \left(\frac{V}{V_0} \right)^{-2/3} \left[1 - \left(\frac{V}{V_0} \right)^{1/3} \right] \exp \left(\frac{3}{2} (K'_0 - 1) \left(1 - \left(\frac{V}{V_0} \right)^{1/3} \right) \right)$$

with $V_0 = 14.52 \text{ cm}^3 \text{ mol}^{-1}$ ($a_0 = 3.64 \text{ Å}$), $K_0 = 4.26 \text{ GPa}$, $K'_0 = 7.75$. The pressure was determined by both the non-hydrostatic ruby¹⁸ and the $\text{SrB}_3\text{O}_7\text{:Sm}^{2+}$ (ref. 19) luminescence pressure scales, and also by the Au molar volume (estimated from the 111 and 002 diffraction peaks) related to the pressure through the Au equation of state²⁰. The X-ray beam was collimated by microslits down to $15 \times 15 \text{ μm}$. correction for the pressure gradient was estimated from the variation of the d -spacings of the strong 110 peak within the H_2O single crystal. The error bar in the pressure measurement increases from around $\pm 2 \text{ GPa}$ at 50 GPa to $\pm 6 \text{ GPa}$ at 170 GPa. The effect of the non-hydrostatic stress leads to a 3% dispersion for the lattice parameter in the 100 GPa range. A stress analysis, based on a simple elastic-type calculus and on the use of the elastic coefficients of ice VII measured up to 7 GPa (ref. 21) and linearly extrapolated with pressure above this, was performed to obtain the correct value of the cell parameter. The value of the uniaxial stress component and of the corresponding hydrostatic lattice parameter were fitted by least squares. A solution could always be found that reproduces the observed d -spacings very well. The difference between the corrected volume and the volume estimated from the mean value of the lattice parameter calculated over the various reflections was $\sim 1\%$ in the 100 GPa range. Hence, half of it (0.5%) should maximize the size of the error bar for the volume determination above 30 GPa.

corrections for the pressure gradient and for the effect of the uniaxial stress have to be made. Extending the measurements to 170 GPa allows the EOS to be determined more accurately.

The single-crystal X-ray diffraction pattern associated with the 222 reflection is presented in Fig. 2, top inset; two other reflections are also clearly seen. The 111 reflection has been observed in neutron diffraction studies¹. (Its intensity change will be analysed below.) In contrast, the superlattice reflection at nearly $\frac{1}{2}\frac{1}{2}\frac{1}{2}$ was a surprise. Its d -spacing is slightly different from the value expected from a doubling of the unit cell, indicating a periodic modulation of the structure along the $[111]$ direction. The evolution with pressure of this peak (which we denote as $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$) was studied for three large single crystals (using 500- μm anvil-tip diameter) under hydrostatic compression. The following observations will support the interpretation proposed below. The $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$ peak could be measured in all the four $[111]$ directions accessible within the X-ray aperture of the diamond anvil cell, with identical d -spacings (within error limits) and very similar intensity. Neither the $\frac{1}{4}\frac{1}{4}\frac{1}{4}$ reflection nor a similar modulation in the $[110]$, $[200]$ or $[121]$ directions could be detected. No distortion from the cubic structure was observed. The periodicity of the modulation, $d_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}/4d_{222}$, is plotted versus pressure in the main part of Fig. 2. The change with pressure is continuous and very reproducible between the three samples up to 8 GPa, independent of the orientation of the crystal with respect to the load axis on the diamond anvils or of the thickness of the gold ring in the sample chamber. Also, the $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$ peak was observed in each run at the pressure when the single crystal started to grow from the melt. These two last observations rule out the possibility that this new peak might be produced by uniaxial stress. We believe that the modulated phases are intrinsic properties of ice VII that tend to disappear in a strained crystal or in powdered ice. Between 10 and 20 GPa, the modulation seems to show lock-in plateaux, as predicted by theories of modulated systems^{9,13}; we note that this corresponds to the pressure range where the $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$ peak has maximum intensity (Fig. 2 inset).

We consider that there is an analogy between the present observations and the phase diagram of the three-dimensional Ising model with competing interactions⁹. First, calculations showed that a large fraction of the phase diagram of the Ising model is formed by the ‘++--++’ phase, consisting of two ‘up’ layers followed by two ‘down’ layers and so on. At low temperature, ice VII transforms into a closely related structure¹, ice VIII, which is

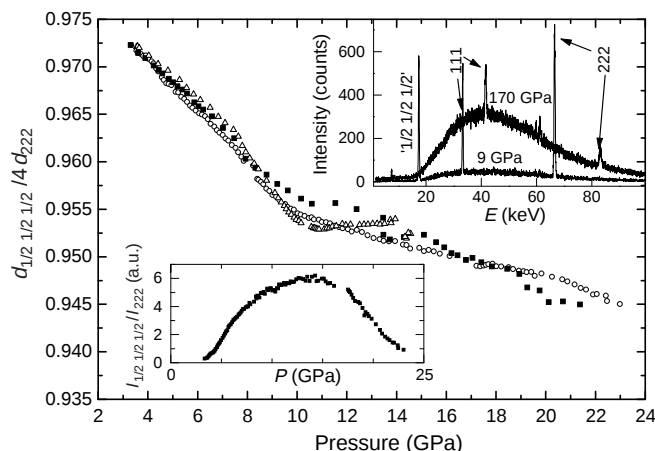


Figure 2 Periodicity of the modulation $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$ versus pressure in ice VII at 300 K. Main figure, variation with pressure of the ratio of the d -spacing of the peak $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$ to its ideal value. Measurements on three different samples are indicated by three different symbols. Top inset, single-crystal energy-dispersive spectra for the 222 reflection at 9 GPa and 170 GPa. The 111 and $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$ peaks are also clearly identified. Bottom inset, evolution with pressure of the intensity of the $S_{\frac{1}{2}\frac{1}{2}\frac{1}{2}}$ reflection relative to that of the 222 reflection (not corrected for the wiggler energy spectrum).

a dipole-ordered antiferroelectric solid. The dipole ordering is accompanied by a slight tetragonal distortion of the cubic unit cell. The 202 peak of ice VIII corresponds to the 111 peak of ice VII, and the allowed 101 peak of ice VIII (weakly measured on powder with X-rays and with a d -spacing exactly twice that of the 202 peak) corresponds to the forbidden $\frac{111}{222}$ peak of ice VII. Hence, neglecting the tetragonal distortion, the arrangement of the dipoles in ice VIII can be seen as a ‘++--++’ order, consisting of two quasi-(111) layers where the dipoles point positively along the [001] direction, followed by two quasi-($\bar{1}\bar{1}\bar{1}$) layers where the dipoles point negatively along the [001] direction. Second, as the ratio of the competing next-nearest-neighbour interaction to the nearest-neighbour interaction is changed, the Ising model displays a sequence of spatially modulated phases; the periodicity of the modulation changes quasi-continuously with the ratio of the interactions. So in ice VII we could be observing an interesting stability field of modulated phases around the ‘++--++’ ice VIII phase, due to the magnitudes of the competing dipolar interactions being changed by pressure. Unravelling the detailed arrangement of hydrogen atoms is beyond the present experimental study, but the [111] directions now appear as the axes around which to construct models of disorder in ice VII.

As seen in Fig. 2, top inset, the 111 peak can be clearly followed up to 170 GPa. The calculation of the structure factor shows that its intensity can be considered as entirely due to the hydrogens, because the contribution of the oxygen atoms (even when taking into account possible oxygen disorder) is at most 1% of the total intensity of the peak. The measured change in the ratio of the intensity of the 111 peak to that of the 222 peak, I_{111}/I_{222} , should be quite accurate in the present single-crystal technique. The evolution of I_{111}/I_{222} with pressure (measured on two [111] directions with good agreement) is shown in Fig. 3. In an isolated-atom approximation of the structure factor, I_{111}/I_{222} should continuously increase by a factor of about 3 as H moves from its position in ice VII towards

the centre of the two oxygens. Trying to model more realistically the O–H covalent bonding and its change with pressure, by placing charges half-way between O and H, has only a small effect—at maximum, a 20% change in the factor 3 of the increase (when $\frac{1}{2}$ an electron is placed at the mid-point of the O–H bond). Taking into account the quantum smearing of the proton due to tunnelling⁵ lessens this ratio by 10% at most. Finally, a plausible way to reach an increase of around 5 for I_{111}/I_{222} is to model the oxygen atoms as disordered over displaced sites. We can therefore confidently state that the evolution of I_{111}/I_{222} from 1 to a plateau at around 3 is the signature of the transition to symmetric ice, ice X. We note that the structure factor calculation, taking into account the experimental inputs of the EOS and of the small change of the O–H bond-length, ($d_{\text{O-H}}$) with pressure¹⁴ measured up to 20 GPa, gives a linear increase to a 50% variation of I_{111}/I_{222} at 20 GPa. We can now interpret the data in Fig. 3. (1) The linear fit through the experimental points up to at least 20 GPa reflects the very small variation of $d_{\text{O-H}}$ in this pressure range^{2,14}. (2) The plateau of I_{111}/I_{222} at a value of about 3 reveals the stability domain of ice X over the pressure range about 60 GPa. This is in very good agreement with the infrared absorption studies^{6,7} that provided evidence of changes in the infrared spectra at ~ 67 GPa, subsequently identified by an *ab initio* calculation as the fingerprints of the stability of ice X (ref. 15). However, the transformation from ice VII to ice X can be more or less sluggish. In our good single-crystal experiment (filled circles), the transition was more sudden and correspondingly the lattice parameter in the pressure range 30–70 GPa falls below the average EOS in Fig. 1. In this case, the transition to ice X occurs as low as about 40 GPa, in agreement with a previous claim from Brillouin spectroscopy⁸. In another case, when the crystal is broken up, the transformation could take place through the formation of a dipolar glass, as proposed previously¹². (3) The rapid rise of I_{111}/I_{222} at pressures greater than 150 GPa could indicate that a displacive phase transition is taking place, such as the recently predicted body-centred-cubic/hexagonal-close packed transformation path for the oxygens with the hydrogens rigidly following this displacement in between two oxygens¹⁶. An onset of oxygen site disordering, associated with a large anharmonicity, could also explain the resonance and intensity change observed in the infrared spectra at around 150 GPa (ref. 7). Unfortunately, the change on the d -spacings was too small to be detected at the beginning of this transformation.

The centring process is visualized in Fig. 3 inset by plotting $d_{\text{O-H}}$ versus the nearest-neighbour O–O distance, $d_{\text{O-O}}$. The stability of ice X at around 65 GPa implies that $d_{\text{O-H}}$ becomes equal to $d_{\text{O-O}}/2$ at that pressure. This point complements the neutron data up to 20 GPa^{2,14}. The results from our experiment are compared with a very rough approximation of $d_{\text{O-H}}$ versus $d_{\text{O-O}}$, calculated from the assumption that the small linear increase of $d_{\text{O-D}}$ with pressure measured¹⁴ up to 20 GPa can be extrapolated to very high pressure; the predictions of the recent quantum simulation⁵ are also shown in this inset. It is clear that the very small variation of $d_{\text{O-H}}$ with pressure measured to 20 GPa is strongly accelerated in approaching ice X, in agreement with an O–H stretching mode softening at the transition⁶. We note that the structural transformation to ice X seems to be accurately described by path-integral *ab initio* simulations⁵, indicating that a detailed understanding of quantum effects in proton transfer systems is within the capabilities of numerical methods. □

Received 27 January; accepted 12 October 1998.

- Kuhs, W. F., Finney, J. L., Vettier, C. & Bliss, D. V. Structure and hydrogen ordering in ices VI, VII and VIII by neutron powder diffraction. *J. Chem. Phys.* **81**, 3612–3623 (1984).
- Nelmes, R. J., Loveday, J. S., Marshall, W. G., Hamel, G., Besson, J. N. & Klotz, S. Multisite disordered structure of ice VII to 20 GPa. *Phys. Rev. Lett.* **81**, 2719–2722 (1998).
- Holzappel, W. B. On the symmetry of the hydrogen bonds in ice VII. *J. Chem. Phys.* **56**, 712–715 (1972).
- Schweizer, K. S. & Stillinger, F. H. High pressure phase transitions and hydrogen-bond symmetry in ice polymorphs. *J. Chem. Phys.* **80**, 1230–1240 (1984).

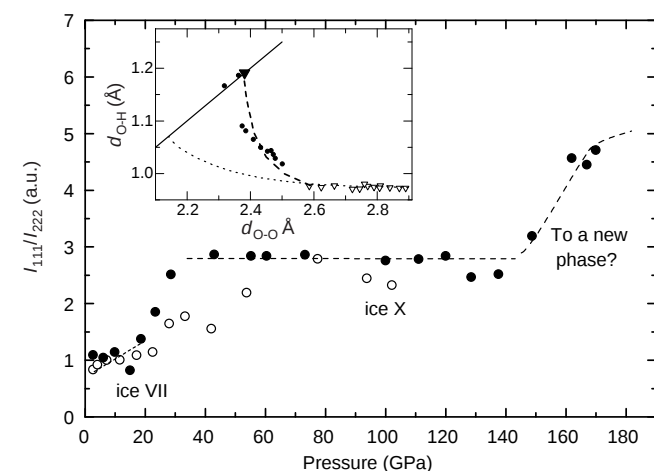


Figure 3 Proton ordering with pressure in ice at 300K from the intensity change of the 111 reflection. Main figure, evolution with pressure of the ratio of the intensity of the 111 reflection to the intensity of the 222 reflection (in arbitrary unit), reflecting changes in proton ordering. The linear dashed line up to 20 GPa represents the atomic structure-factor calculation with experimental $d_{\text{O-H}}$ (refs 2, 14). The plateau at ~ 3 is the signature of ice X. Inset, evolution of $d_{\text{O-H}}$ versus $d_{\text{O-O}}$. The triangles are the neutron data^{2,14} (because tunnelling is not expected to play a role in this pressure range, the O–H bond-length should not be significantly different). The filled triangle is the data deduced from the volume determination at the centring pressure, 65 GPa. The filled circles correspond to the quantum simulations⁵. The dotted line represents the simple empirical model based on the extrapolation of the constant linear variation of $d_{\text{O-H}}$. The dashed line is a reasonable interpolation between experimental data. The symmetrization line, $d_{\text{O-H}} = d_{\text{O-O}}/2$, is indicated (solid line).

5. Benoit, M., Marx, D. & Parrinello, M. Tunnelling and zero-point motion in high-pressure ice. *Nature* **392**, 258–261 (1998).
6. Aoki, K., Yamawaki, H., Sakashita, M. & Fujihisa, H. Infrared absorption study of the hydrogen-bond symmetrization in ice to 100 GPa. *Phys. Rev. B* **54**, 15673–15677 (1996).
7. Goncharov, A. F., Struzhkin, V. V., Somayazulu, M. S., Hemley, R. J. & Mao, H. K. Compression of ice to 210 GPa: Infrared evidence for a symmetric hydrogen-bonded phase. *Science* **273**, 218–220 (1996).
8. Polian, A. & Grimsditch, M. New high pressure phase of H₂O: Ice X. *Phys. Rev. Lett.* **52**, 1312–1314 (1984).
9. Bak, P. & von Boehm, J. Ising model with solitons, phasons and 'the devil's staircase'. *Phys. Rev. B* **21**, 5297–5308 (1980).
10. Loubeyre, P. *et al.* X-ray diffraction and equation of state of hydrogen at megabar pressures. *Nature* **383**, 702–704 (1996).
11. Hemley, R. J. *et al.* Static compression of H₂O-ice to 128 GPa. *Nature* **330**, 737–740 (1987).
12. Wolanin, E. *et al.* Equation of state of ice VII up to 106 GPa. *Phys. Rev. B* **56**, 5781–5785 (1997).
13. Aubry, S. Devil's staircase and order without periodicity in classical condensed matter. *J. Phys. (Paris)* **44**, 147–162 (1983).
14. Nemes, R. J. *et al.* Neutron diffraction study of the structure of deuterated ice VIII to 10 GPa. *Phys. Rev. Lett.* **71**, 1192–1195 (1993).
15. Bernasconi, M., Silvestrelli, P. L. & Parrinello, M. Ab-initio infrared absorption study of the hydrogen-bond symmetrization in ice. *Phys. Rev. Lett.* **81**, 1235–1238 (1998).
16. Benoit, M., Bernasconi, M., Focher, P. & Parrinello, M. New high pressure phase of ice. *Phys. Rev. Lett.* **76**, 2934–2936 (1996).
17. Vinet, P., Ferrante, J., Smith, J. R. & Rose, J. H. A universal equation of state for solids. *J. Phys. C* **19**, L467–L473 (1986).
18. Mao, H. K., Bell, P. M., Shaner, J. & Steinberg, D. Specific volume measurements of Cu, Mo, Pd and Ag and calibration of the ruby R1 fluorescence pressure gauge from 0.06 to 1 Mbar. *J. Appl. Phys.* **49**, 3276–3283 (1978).
19. Datchi, F., LeToullec, R. & Loubeyre, P. Improved calibration of the SrB₄O₇:Sm²⁺ optical pressure gauge: Advantages at very high pressures and high temperatures. *J. Appl. Phys.* **81**, 3333–3339 (1997).
20. Anderson, O. L., Isaak, O. & Yamamoto, S. Anharmonicity and the equation of state of gold. *J. Appl. Phys.* **65**, 1534–1543 (1989).
21. Shimizu, H., Nabetani, T., Nishiba, T. & Sasaki, S. High pressure elastic properties of the VI and VII phase of ice in dense H₂O and D₂O. *Phys. Rev. B* **53**, 6107–6110 (1996).

Acknowledgements. We thank J. Loveday for the computation of the effect of oxygen disorder and proton tunnelling on the structure factor, F. Datchi for help in the experiments and the ESRF staff for technical support. We also thank Y. Petroff for continuous interest in this work, and Ph. Pruzan and J. M. Besson for discussions. We acknowledge a referee for suggesting the interpretation of the rise in I_{111}/I_{222} at 150 GPa.

Correspondence and requests for materials should be addressed to P.L. (paul.loubeyre@bruyeres cea.fr).

Circularly polarized light generated by photoexcitation of luminophores in glassy liquid-crystal films

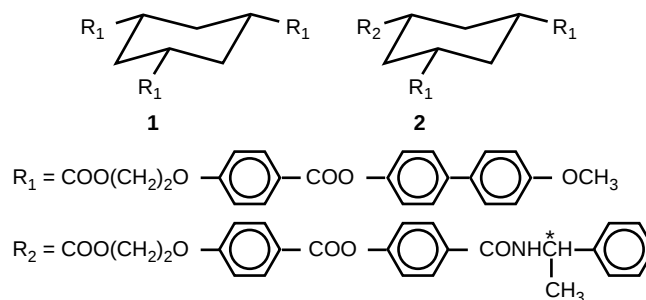
S. H. Chen^{†‡§}, D. Katsis[†], A. W. Schmid[‡], J. C. Mastrangelo[§], T. Tsutsui^{||} & T. N. Blanton[¶]

^{*} Materials Science Program, [†] Department of Chemical Engineering, [‡] Laboratory for Laser Energetics, [§] NSF Center for Photoinduced Charge Transfer, Center for Optoelectronics and Imaging, University of Rochester, 240 East River Road, Rochester, New York 14623-1212, USA
^{||} Department of Applied Science for Electronics and Materials, Kyushu University, Kasuga, Fukuoka 816-8580, Japan
[¶] Analytical Technology Division, Kodak Research Laboratories B82A, Eastman Kodak Company, Rochester, New York 14650-2106, USA

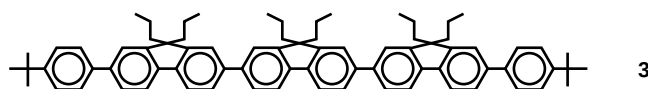
Optical information processing, display and storage can be accomplished with linearly or circularly polarized light. In passive (non-emitting) devices, linear polarization can be produced by anisotropic absorption of light¹, whereas circular polarization has been attained by selective reflection of unpolarized light propagating through a chiral-nematic liquid-crystal film². Active (light-emitting) devices capable of polarized emission are also needed. In principle, optical and electronic excitation of materials containing uniaxially and helically arranged luminophores should produce linearly and circularly polarized emission, respectively. In practice, the former is easier to achieve and is therefore more technologically advanced^{3–8}. Here we report the generation of strongly circularly polarized photoluminescence from films of glass-forming chiral-nematic liquid crystals⁹ in which are embedded light-emitting dopants. This host material apparently induced alignment of the luminophores to a degree that produces almost pure circular polarization within the 400–420-nm wave-

length band of the emitted light. We anticipate that composite films of this sort might find applications within photonic technology such as colour-image projection¹⁰ and stereoscopic displays¹¹.

As a measure of the degree of circularly polarized photoluminescence (CPPL), the dissymmetry factor, g_c , is defined as $2(I_L - I_R)/(I_L + I_R)$, where I_R and I_L denote respectively the right- and left-handed emission intensity. It is evident that $|g_c|$ equals 2 for pure, single-handed CPPL. Polythiophene with chiral pendants in solution¹² and chiral-nematic fluid films doped with a fluorescent dye¹³ have been found to produce a g_c value of 5.0×10^{-3} and 0.8, respectively. Circularly polarized electroluminescence has also been demonstrated using a thin film of poly(*p*-phenylene vinylene) with chiral pendants, yielding a g_c value of 1.7×10^{-3} (ref. 14). All these previous studies have been restricted to emission outside the selective reflection band (that is, outside the resonance region). Here we study photoluminescence inside the resonance region. A nematic and a chiral-nematic glass-forming liquid crystal (GLC) depicted respectively as **1** and **2** below, were synthesized as part of our recent work on GLCs as an emerging class of optical materials⁹:



As a light-emitting dopant, Exalite 428 (Exciton, Inc., Dayton, Ohio, USA) with the following structure was used:



Compound **3** has an absorption peak at 360 nm with an extinction coefficient of $1.30 \times 10^5 \text{ M}^{-1} \text{ cm}^{-1}$, and an emission peak covering 380 to 500 nm with excitation at 360 nm; these properties were obtained from a 10^{-5} M solution of **3** in methylene chloride. A series of chiral-nematic hosts with a wide range of selective reflection wavelength (λ_R) were prepared by dissolving various ratios of **1** and **2** in methylene chloride. Doping of the chiral-nematic hosts with Exalite 428 at a level of 0.2 wt% was also accomplished by co-dissolution: solid samples were collected by evaporating off the solvent *in vacuo*. The thermotropic properties and the measured values of λ_R of the GLC hosts, (I) to (V), are summarized in Table 1.

Table 1 Chemical composition and properties of GLC hosts

GLC host	1 (wt%)	2 (wt%)	DSC* (°C)	$\lambda_R^\dagger$ (nm)
(I)	0	100	G 77 Ch 147 I	410
(II)	5	95	G 75 Ch 154 I	434
(III)	49	51	G 68 Ch 215 I	890
(IV)	62	38	G 67 Ch 235 I	1210
(V)	100	0	G 64, 95 K ₁ 130 K ₂ 180 N 279 I†	α

Formulae of compounds **1** and **2** are given in the text.

* Thermal transition temperatures were determined from DSC second heating scans at $50^\circ \text{C min}^{-1}$ of samples preheated to 250°C followed by quenching at $-60^\circ \text{C min}^{-1}$ to -30°C ; the nematic and cholesteric mesophases were identified as threaded textures and oily streaks, respectively, under polarizing optical microscopy. Symbols: G, glassy; K, crystalline; Ch, cholesteric; N, nematic; I, isotropic.

† Selective reflection wavelength, λ_R , was determined for 14- μm -thick film by ultraviolet-visible spectrophotometry.

‡ Recrystallization at 130°C and crystalline modification at 180°C were observed in the DSC heating scan of vitrified (V) at $50^\circ \text{C min}^{-1}$.

The differential scanning calorimetry (DSC) data indicate that all the hosts are capable of glass formation upon quenching from the liquid-crystalline or isotropic melts. As revealed by the X-ray diffraction (XRD) patterns shown in Fig. 1, pristine samples of (I) and (V) are polycrystalline (traces a and c), while thermally processed films are glassy (traces b and d) with the frozen cholesteric and nematic mesophase identified as oily streaks and threaded textures, respectively, under polarizing optical microscopy. The thermal transition temperatures and λ_R values reported in Table 1 for the GLC hosts remained unchanged on doping with Exalite 428 at 0.2 wt%.

One objective of this study was to examine the effect on CPPL of the structural periodicity imposed on the luminophore. Thin films were prepared between two alignment-coated, fused silica substrates using glass-fibre spacers for thickness control. The devices were heated to 210 °C to melt the crystalline powder, then cooled to 130 °C for shearing to induce alignment followed by 3 minutes of annealing before quenching to room temperature. The optical set-up depicted in Fig. 1 of ref. 15 was used for collecting CPPL spectra. To validate the capability for CPPL analysis, an isotropic film of poly(methyl methacrylate) (Polysciences, Inc., Warrington, Pennsylvania, USA) doped with Exalite 428 at 0.2 wt% was found to yield a g_e value of zero across the spectral region from 380 to 600 nm. The CPPL spectra of films (I) and (II), accompanied by the selective reflection bands, are shown in Fig. 2, where the cross-over between I_R and I_L is revealed. For films (III) and (IV), CPPL is characterized by $I_R > I_L$ in the entire spectral region because of the left-handed helical structure of the host. The g_e values calculated from the CPPL spectra are plotted in Fig. 3a for all four chiral-nematic films having the same thickness, τ , of 14 μm . The nearly constant g_e values with respect to wavelength found for

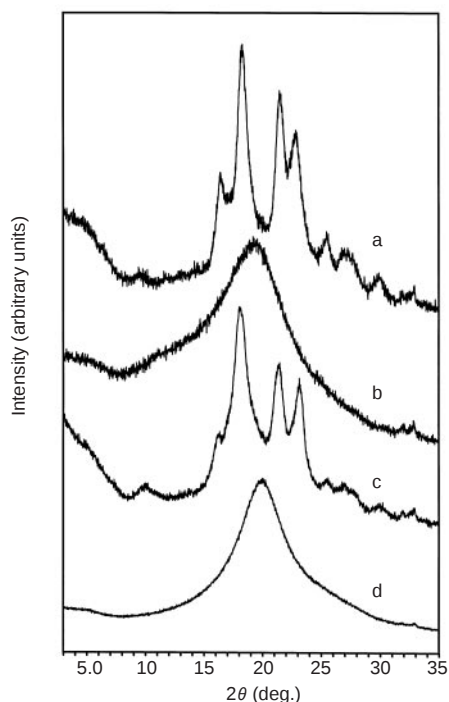


Figure 1 X-ray diffraction patterns of the glassy liquid-crystalline hosts. Trace a, pristine powder of (I) at 26 °C; trace b, (I) heated to 220 °C on a platinum strip then quenched at -60 °C min^{-1} to 26 °C; trace c, pristine powder of (V) at 26 °C; trace d, (V) heated on a platinum strip to 220 °C then quenched at -60 °C min^{-1} to 26 °C. Scans a and c reveal the polycrystalline nature of pristine samples, while scans b and d confirm the glassy characteristic of the thermally processed, chiral-nematic and nematic films. Left at 26 °C for over 3 months, the thermally processed films did not recrystallize, as evidenced by X-ray diffraction. Moreover, the optical clarity of the films sandwiched between fused silica substrates for CPPL measurements did not deteriorate when left at 26 °C over a period of 12–15 months.

films (III) and (IV) are characteristic of photoluminescence outside the resonance region.

To characterize the films, we employed recent CPPL theory¹³ to determine the second- and fourth-rank order parameters, $\langle P_2 \rangle$ (that is, S_{em}) and $\langle P_4 \rangle$, governing emission from quasi-nematic layers comprising the chiral-nematic film. We assumed that the absorption and emission transition dipole moments are both parallel to the long molecular axis of Exalite 428; we also assumed that $S_{ab} = S_{em} = S$, where S_{ab} is the order parameter governing anisotropic absorption at quasi-nematic layers. Input data to the CPPL theory include the film thickness (14 μm), excitation wavelength (370 nm), emission wavelength (428 nm), average refractive index, $\bar{n}$, optical birefringence, Δn , and the helical pitch length, $p = \lambda_R/\bar{n}$. Another piece of information needed is the absorption coefficient, A/τ , in which A is the average absorbance evaluated as follows: $A = (A_{\parallel} + 2A_{\perp})/3$, where $A_{\parallel}$ and $A_{\perp}$ are the absorbance measured parallel and perpendicular, respectively, to the nematic director. A nematic film 14 μm thick was prepared with (V), doped with Exalite 428 at 0.2 wt%; the sample was heated to 250 °C, at which temperature annealing was allowed for 1 hour, followed by quenching to room temperature. Ultraviolet–visible absorption measurements led to $A/\tau = 154\text{ cm}^{-1}$. Linearly polarized photoluminescence was characterized for the nematic film to yield $S = 0.60$ and $\langle P_4 \rangle = -0.022$, using equations (39) and (40) in ref. 13. Refractometric measurement at 589 nm resulted in $\Delta n = 0.23$ and $\bar{n} = 1.67$, which were used in the determination of Δn and $\bar{n}$ as functions of wavelength by applying a phase-difference technique¹⁶ to film (V). It was found that $\Delta n = 0.35$ and $\bar{n} = 1.70$ at 370 nm, and $\Delta n = 0.29$ and $\bar{n} = 1.67$ at 428 nm. With Δn , $\bar{n}$, and A/τ having been evaluated, the best fit to the theory of the observed g_e at 428 nm yielded $S = 0.55$ and $\langle P_4 \rangle = 0.065$ for film (III), and $S = 0.66$ and $\langle P_4 \rangle = 0.020$ for film (IV). Therefore, all the chiral-nematic films, including (I) and (II), prepared following the same procedures are expected to have an average order parameter of $S = 0.61 \pm 0.05$.

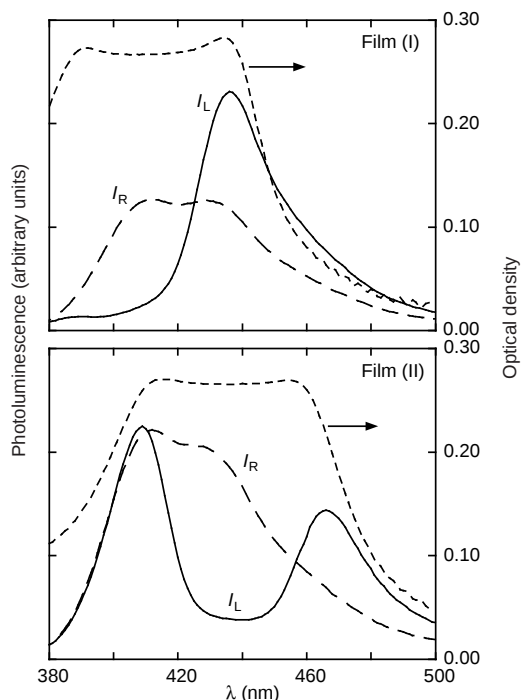


Figure 2 Circularly polarized photoluminescence spectra. The solid and long-dashed curves show I_L and I_R respectively. Also shown are selective reflection bands (short-dashed curves) for 14- μm -thick films of (I) and (II), using an unpolarized light source centred at 370 nm with a full-width at half-maximum of 15 nm. The measurements were performed with excitation and emission beams normal to the film surface in a straight-through arrangement.

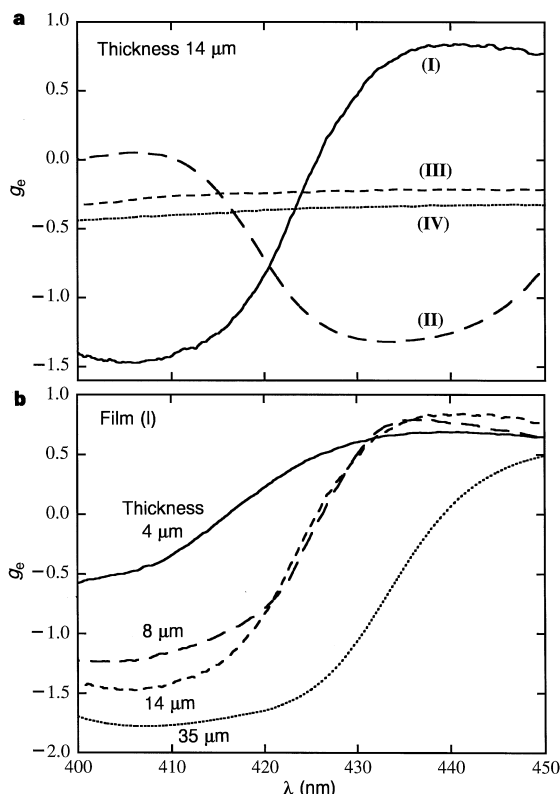


Figure 3 Dissymmetry factor (g_e) calculated from the CPPL spectra. **a**, Films (I), (II), (III) and (IV), all 14 μm thick; **b**, effect of film thickness on g_e displayed for (I). The reported g_e curves are reproducible to within $\pm 5\%$ in general.

As shown in Fig. 3a, films (III) and (IV) produced a small, negative g_e that remains largely constant throughout the spectral region, supporting the assumption that the emission of Exalite 428 is characterized by a single dipole moment parallel to the long molecular axis. In sharp contrast to the typical CPPL behaviour outside the resonance region, films (I) and (II) show emission in the resonance region, resulting in a large g_e that undergoes handedness reversal in a single-handed chiral-nematic host. This is an observation inexplicable by CPPL theory, in contrast to emission outside the resonance region^{13,15}. The effect of thickness on g_e for film (I) is assessed in Fig. 3b, revealing the general cross-over behaviour for all thicknesses. We found nearly pure right-handed photoluminescence, manifested as g_e approaching -2 , in the 400–420 nm wavelength range at a thickness of 35 μm . In addition, the single-photon counting technique was employed to determine the excited-state lifetime, t_{ex} . For emission outside the resonance region, a constant $t_{\text{ex}} = 0.66 \pm 0.02$ ns across the emission spectrum was found. For emission inside the resonance region, we found $t_{\text{ex}} = 0.67 \pm 0.02$ ns. The lack of a prolonged t_{ex} inside the resonance region alone seems to suggest the absence of microcavity processes.

Well aligned solid films with superior morphological stability hold promise for practical applications. Specific examples include colour image projection¹⁰ and stereoscopic displays¹¹ that have been proposed on the basis of non-emitting chiral-nematic films. Furthermore, embedding laser dyes in vitrified chiral-nematic films has the potential to realize low-threshold distributed feedback lasers¹⁷, with a cost advantage over inorganic semiconductor lasers.

Received 11 June; accepted 30 November 1998.

- Dirix, Y., Tervoort, T. A. & Bastiaansen, C. Optical properties of oriented polymer/dye polarizers. *Macromolecules* **28**, 486–491 (1995).
- Schadt, M. & Fünfschilling, J. New liquid crystal polarized color projection principle. *Jpn J. Appl. Phys.* **29**, 1974–1984 (1990).

- Montali, A., Bastiaansen, C., Smith, P. & Weder, C. Polarizing energy transfer in photoluminescent materials for display applications. *Nature* **392**, 261–264 (1998).
- Dyreklev, P. et al. Polarized electroluminescence from an oriented substituted polythiophene in a light emitting diode. *Adv. Mater.* **7**, 43–45 (1995).
- Era, M., Tsutsui, T. & Saito, S. Polarized electroluminescence from oriented *p*-sexiphenyl vacuum-deposited film. *Appl. Phys. Lett.* **67**, 2436–2438 (1995).
- Cimrová, V., Remmers, M., Neher, D. & Wegner, G. Polarized light emission from LEDs prepared by the Langmuir-Blodgett technique. *Adv. Mater.* **8**, 146–149 (1996).
- Sariciftci, N. S., Lemmer, U., Vacar, D., Heeger, A. J. & Janssen, R. A. J. Polarized photoluminescence of oligothiophenes in nematic liquid crystalline matrices. *Adv. Mater.* **8**, 651–654 (1996).
- Lüssem, G. et al. Polarized electroluminescence of light emitting liquid crystalline polymers. *Liq. Cryst.* **21**, 903–907 (1996).
- Chen, S. H., Shi, H., Conger, B. M., Mastrangelo, J. C. & Tsutsui, T. Novel vitrifiable liquid crystals as optical materials. *Adv. Mater.* **8**, 998–1001 (1996).
- Heyndericks, I. E. J. R. & Broer, D. J. Illumination system for a color projection device and circular polarizer suitable for use in such an illumination system, and color image projection device comprising such an illumination system and circular polarizer. US Patent No. 5,626,408 (1997).
- Hall, D. R. Use of stereoscopic systems utilizing chiral liquid crystals. US Patent No. 5,699,184 (1997).
- Langeveld-Voss, B. M. W. et al. Circular dichroism and circular polarization of photoluminescence of highly ordered poly[3,4-di[(S)-2-methylbutoxy]-thiophene]. *J. Am. Chem. Soc.* **118**, 4908–4909 (1996).
- Shi, H., Conger, B. M., Katsis, D. & Chen, S. H. Circularly polarized fluorescence from chiral nematic liquid crystalline films: theory and experiment. *Liq. Cryst.* **24**, 163–172 (1998).
- Peeters, E. et al. Circularly polarized electroluminescence from a polymer light-emitting diode. *J. Am. Chem. Soc.* **119**, 9909–9910 (1997).
- Katsis, D., Schmid, A. W. & Chen, S. H. Mechanistic insight into circularly polarized photoluminescence from a chiral-nematic film. *Liq. Cryst.* (in the press).
- Wu, S. T., Efron, U. & Hess, L. D. Birefringence measurements of liquid crystals. *Appl. Opt.* **23**, 3911–3915 (1984).
- Il'chishin, I. P., Tikhonov, E. A., Tishchenko, V. G. & Shpak, M. T. Generation of a tunable radiation by impurity cholesteric liquid crystals. *JETP Lett.* **32**, 24–27 (1981).

Acknowledgements. We thank R. W. Boyd, L. J. Rothberg and S. D. Jacobs for discussions. This work was supported by the US National Science Foundation, US Department of Energy, and the Japanese Ministry of International Trade and Industry.

Correspondence and requests for materials should be addressed to S.H.C. at the University of Rochester (e-mail: shch@lle.rochester.edu).

Accelerated dissolution of diatom silica by marine bacterial assemblages

Kay D. Bidle & Farooq Azam

Marine Biology Research Division, 0202, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA

Downward fluxes of biogenic silica and organic matter in the global ocean derive dominantly from the productivity of diatoms—phytoplankton with cell walls containing silica encased in an organic matrix^{1,2}. As diatoms have an absolute requirement for silicon (as silicic acid)³, its supply into the photic zone—largely by silica dissolution and upwelling—controls diatom production (and consequently the biological uptake of atmospheric CO₂ by the ocean) over vast oceanic areas⁴. Current biogeochemical models assume silica dissolution to be controlled by temperature, zooplankton grazing and diatom aggregation^{4,5}, but the role of bacteria has not been established. Yet bacteria utilize about half of the organic matter derived from oceanic primary production⁶ by varied strategies, including attack on dead and living diatoms by using hydrolytic enzymes^{7,8}, and could adventitiously hasten silica dissolution by degrading the organic matrix which protects diatom frustules from dissolution^{9,10}. Here we report the results of experiments in which natural assemblages of marine bacteria dramatically increased silica dissolution from two species of lysed marine diatoms compared to bacteria-free controls. Silica dissolution accompanied, and was caused by, bacterial colonization and hydrolytic attack. Bacteria-mediated silicon regeneration rates varied with diatom type and bacterial assemblage; observed rates could explain most of the reported upper-ocean silicon regeneration^{5,11}. Bacteria-mediated silicon regeneration may thus critically control diatom productivity and the cycling and fate of silicon and carbon in the ocean.

In the recent concept of the “silicate pump”^{3,4}, diatom productivity

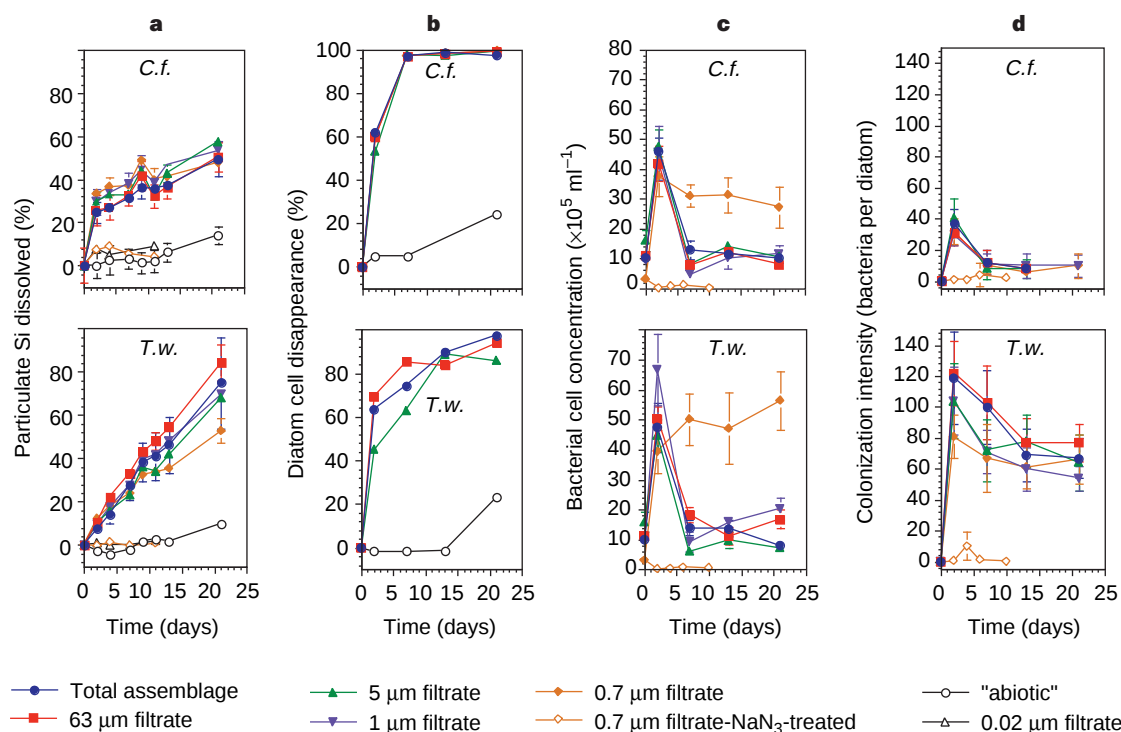


Figure 1 Bacterial colonization and silicon regeneration of diatoms lysed by freeze-thaw. (*C.f.*, *Cylindrotheca fusiformis*; *T.w.*, *Thalassiosira weissflogii*.) Diatom detritus was exposed to whole sea water, sea water filtered through various size filters, or 0.2- μm -filtered, autoclaved sea water ("abiotic") (see symbol key); additional samples were treated with sodium azide (50 mM) or

suspended in 0.02- μm -filtered sea water. **a**, Dissolution of particulate silica; **b**, "disappearance" of diatoms (no longer visible in the epifluorescence microscope by their chlorophyll autofluorescence or other cell components); **c**, bulk-phase bacterial concentrations; and **d**, bacterial colonization intensity.

in open-ocean upwelling systems becomes silicon-limited because of selective recycling of nitrogen over silicon: zooplankton grazing on diatoms efficiently regenerates nitrogen within the upper mixed layer but packages silica in faecal pellets, transporting it to depth where it dissolves to silicic acid. Consequently, silicic acid supply rate (rather than iron or nitrate) by upwelling sets the upper limit on diatom productivity and export production. On the other hand, substantial and highly variable silicon regeneration has been observed within the water column in several oceanic systems^{5,11}. The global-mean biogenic silica dissolution: production (D:P) ratio is 0.58 (0.05–5.8; ref. 5) with silicon regeneration supporting much (>70%) of the biogenic silica production (diatom growth) in a coastal upwelling area¹², within a warm core ring¹³ and in the Ross Sea continental shelf waters¹. These contradictory views of silicon regeneration raise questions about its mechanism and regulation in the upper mixed layer. Silica dissolution from diatom frustules has been considered as being mainly due to chemical depolymerization¹⁰ not involving biological activity, as metazoan grazing did not enhance regeneration^{14,15}. However, the role of bacteria and the microbial loop in silicon regeneration has not been explicitly examined. Significant silicon regeneration by bacteria would cause a 'leak' in the silica pump, accounting for the high and variable D:P ratios. Pelagic marine bacteria colonize fresh diatom detritus¹⁶, living diatoms⁸ and "marine snow" phytoaggregates⁷ with colonizers expressing very high levels of hydrolytic ectoenzymes (protease, α - and β -glucosidase, alkaline phosphatase, lipase and chitinase). Bacterial hydrolytic attack on diatom cell walls would denude silica of its organic coating and hasten its dissolution.

We tested whether the action of natural, pelagic marine bacterial assemblages on lysing diatoms could accelerate silica dissolution. Our focus was on lysed diatoms because a substantial fraction of diatom productivity, especially after blooms, can escape grazing¹⁷

and instead undergoes cell lysis in response to physiological stress^{18,19}. We used freeze-thaw-lysed axenic cultures of diatom cells; they provided the well-characterized diatom detritus of known history necessary for addressing the mechanistic relationships between bacterial action on diatom detritus and silicon regeneration. Two diatoms were studied, one lightly silicified (*Cylindrotheca fusiformis*; ~ 60 fmol Si per cell) and one heavily silicified (*Thalassiosira weissflogii*; ~ 800 fmol Si per cell). We were mainly interested in regeneration kinetics during the first ~ 10 d, a conservative estimate of the residence time of unaggregated diatom detritus in the upper mixed layer. Lysed diatoms were suspended in whole sea water or in sea water which had been filtered through various pore sizes: 0.2- μm -filtered-autoclaved ("abiotic"), 0.02- μm Whatman Anodisc, GF/F 0.7 μm ("bacterial fraction"), 1- μm Nuclepore, and 5- μm and 63- μm Nitex. This was done to test the contribution of bacterial action to silicon regeneration relative to other trophic levels (for example, protozoa). The added diatom detritus ($(1-3) \times 10^4$ ml⁻¹) and the resulting silica concentrations ($(2-6) \mu\text{g Si l}^{-1}$) were comparable to those during a moderate bloom. We followed silica dissolution and its relationship to microbial populations, particularly those colonizing the detritus, during incubations which were conducted at near-ambient temperature (16–18 °C). Seawater pH (7.9–8.1) did not change during incubation.

When suspended in "abiotic" sea water, the silica of both diatoms was remarkably resistant to dissolution (*C. fusiformis*, $0.2 \pm 0.3\%$ d⁻¹; *T. weissflogii*, $0.3 \pm 0.2\%$ d⁻¹; Fig. 1a). Similarly, low rates (*C. fusiformis*, $0.9 \pm 0.3\%$ d⁻¹; *T. weissflogii*, $0.2 \pm 0.1\%$ d⁻¹; Fig. 1a) were found for diatoms incubated in filtered (0.02- μm pore size), unautoclaved sea water, thus arguing against the involvement of any *in situ* dissolved hydrolases present in sea water or released during filtration. Dissolution was dramatically faster in 0.7- μm -filtered sea water (Fig. 1a), with initial rates

Table 1 Particulate silica dissolution and turnover rates

Diatom	Experiment	Particulate silica dissolved (%) [*]	Turnover rate (% d ⁻¹) [†]	
			Day 0–2	Day >2 [‡]
<i>C. fusiformis</i>	1	81 ± 6.9	4.6 ± 2.8	2.2 ± 0.8
	2	70 ± 1.5	8.2 ± 0.5§	1.5 ± 1.5§
	3	48 ± 2.7	18 ± 1.7	0.8 ± 4.1
	4	52 ± 3.2	17 ± 1.2	0.8 ± 3.2
<i>T. weissflogii</i>	1	53 ± 4.5	2.1 ± 1.2	2.0 ± 0.9
	2	71 ± 0.9	4.0 ± 0.1	2.3 ± 0.8§
	3	55 ± 5.7	6.1 ± 0.7	2.7 ± 1.2
	4	62 ± 4.0	7.9 ± 0.6	3.2 ± 0.2

Data shown are for diatom detritus exposed to natural bacterial assemblages (<0.7 µm).

^{*} At end of experiment (21–31 days); values are shown ± s.d.

[†] Values are shown ± s.d.

[‡] Average turnover rate.

^{||} Water for experiments 3 and 4 was collected at the same time.

[§] Day 0–3.

of $18 \pm 1.7\% \text{ d}^{-1}$ for *C. fusiformis* and $7.9 \pm 0.7\% \text{ d}^{-1}$ for *T. weissflogii*. Regeneration rates were not significantly different when diatom detritus was suspended in the larger size fractions of sea water or in whole sea water (Fig. 1a). Time courses of silica regeneration for the two diatoms were distinctly different, perhaps reflecting differences in the degree of cell-wall silicification as well as cell-wall architecture and chemistry. Turnover rates for the highly silicified *T. weissflogii* were initially lower but later exceeded the *C. fusiformis* rates (Table 1). Consistent with rapid silica dissolution, epifluorescence microscopy showed that lysed diatoms 'disappeared' (sometimes visible as 'ghosts') roughly in parallel with, or faster than, silica dissolution (Fig. 1b). Using epifluorescence and dark-field microscopy, no protozoa were detected in the <0.7-µm fraction nor were any bacteria detected in the <0.02-µm or "abiotic" fractions (except in the 21-d sample for the latter; and this was accompanied by increased silica dissolution). Thus, the <0.02-µm and <0.2-µm incubations were indeed abiotic and the only known biological activity acting on diatom detritus in <0.7-µm incubations would be due to Bacteria (and possibly Archaea²⁰).

In addition to the size-fractionation data, mentioned above, the following lines of evidence suggest that silicon regeneration was indeed caused by bacteria and that the mechanism involved detritus colonization and hydrolytic attack on organic matrix. Sodium azide (50 mM) either strongly inhibited silicon regeneration or abolished it completely (Fig. 1a), roughly in parallel with inhibition of bacterial colonization of detritus (Fig. 1d) and activity. A mixture of penicillin, streptomycin and chloramphenicol, which should have inhibited the activities of many prokaryotes (it inhibited but did not abolish prokaryotic growth in <0.7-µm incubations) also inhibited silica dissolution by 50% after 7–11 d for *C. fusiformis* but no effect was observed for *T. weissflogii*. All biotic incubations showed strong and rapid colonization of diatom detritus by bacteria (bacterial concentration was 10^{11} ml^{-1} in the detritus microenvir-

onment, 10^5 times the bulk-phase concentration; Table 2). Colonization intensity, rather than bulk-phase bacterial abundance, appeared to control silicon regeneration. Colonization intensity and silicon regeneration rates were similar among all biotic incubations, even when bulk-phase bacteria concentrations were very different (for example, <0.7-µm incubations; Fig. 1c); thus, variation in bulk-phase bacterial concentrations and activity may not control silica regeneration. Further, the distinctly different patterns of persistence of colonization for *C. fusiformis* and *T. weissflogii* (Fig. 1d) corresponded to distinctly different time courses of silicon regeneration (Fig. 1a). The colonizing bacteria had high growth rates and imparted extremely high protease activities to the diatom detritus microenvironment (Table 2). Glycoproteins are significant components of diatom cell walls and are thought to be involved in frustule biogenesis²¹ so their hydrolysis by proteases may have hastened silica solubilization. To test that protease action was important in organic matrix dissolution and silica dissolution we incubated uniformly ¹⁴C-labelled diatom detritus with pronase E (0.1 unit ml⁻¹) in the absence of bacteria. Pronase E rapidly released ¹⁴C as well as regenerated silicon at moderate rates (Fig. 2). The pattern of silica dissolution for each diatom following pronase E treatment was similar to that in bacterial treatments (Fig. 1a). Pronase E digestion also yielded distinctly different patterns for the two diatoms in the regeneration of particulate carbon and silicon, presumably reflecting differences in cell-wall structure and perhaps chemical composition.

Rates of bacteria-mediated silica dissolution (Table 1; day 0–2;) were as fast as those reported for acid-digested frustules at a slightly higher temperature ($5.5\text{--}18\% \text{ d}^{-1}$; ref. 9), suggesting that bacteria can efficiently remove the organic matrix. If we assume an average sinking rate of 1 m d^{-1} (ref. 22) and a 20-m mixed-layer depth, then bacteria-mediated dissolution would regenerate 49–51% and 40–73% of silicon from *C. fusiformis* and *T. weissflogii* detritus, respectively (Table 1). These values are comparable to the vertically integrated global-average silica D:P ratio of 0.58 (ref. 5) consistent with an important role for bacteria in silicon regeneration in the sea. Clearly, the *in situ* regeneration rates will be modulated by other ecosystem parameters, particularly those affecting bacterial activity on diatoms. Temperature is commonly considered the main control on diatom silica dissolution²³. Our finding of bacterial mediation of dissolution suggests that temperature controls not only the chemical depolymerization of silica¹⁰ but also the action of bacteria on the organic matrix. Its effect on these two processes is probably sequential and differential, and needs to be elucidated to understand the behaviour of diatom frustules experiencing changing temperature (for example, during sinking). Temperature effects on bacterial metabolism may account for diatom frustules having dissolution coefficients in sea water with a temperature coefficient, Q_{10} , of 2.27 (ref. 9; the study used diatoms collected in net-tows and incubated in 0.45-µm-filtered sea water and probably contained active bacteria). Bacterial species composition and/or metabolic

Table 2 Bacterial colonization and activity on diatom cell detritus

Diatom	Time (d)	Colonization intensity (bacteria cell ⁻¹)	Attached bacteria [†] ($\times 10^{12}$)	Bacteria concentration ratio (Att./FL) [*]	Cell-specific protease hydrolysis activity (amol cell ⁻¹ h ⁻¹)	Protease activity ratio [†] (Att./FL)	Cell-specific growth rate (d ⁻¹)
<i>C. fusiformis</i>	0	–	–	–	17 ± 1.8 [‡]	–	0.04 ± 0.01 [‡]
	1	12 ± 6	40	28,500	84 ± 65	12,900	1.78 ± 1.42
	2	56 ± 14	190	17,500	2,505 ± 1,680	39,100	0.85 ± 1.52
	4	38 ± 16	120	7,200	647 ± 566	6,400	1.66 ± 3.85
<i>T. weissflogii</i>	0	–	–	–	29 ± 4.7 [‡]	–	0.05 ± 0.01 [‡]
	1	16 ± 12	40	27,900	324 ± 274	26,500	2.27 ± 1.92
	2	108 ± 38	270	23,200	4,058 ± 2,698	78,500	1.33 ± 0.95
	4	98 ± 30	250	9,200	11,217 ± 18,781	155,000	6.65 ± 11.5

See Methods for details of experimental procedure.

^{*} Att., attached bacteria; FL, free-living bacteria.

[†] Activities were calculated per unit volume (ml); see Methods.

[‡] Day 0 values are for bulk sea water.

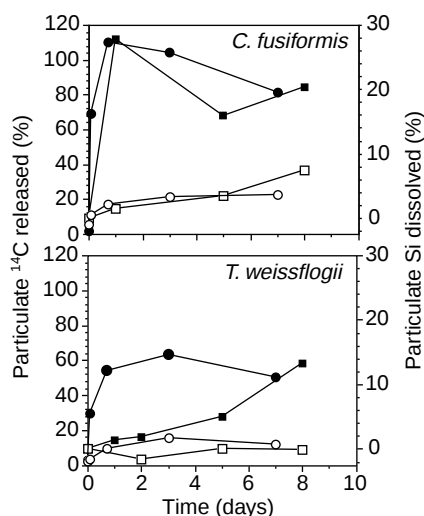


Figure 2 Pronase E digestion of *C. fusiformis* or *T. weissflogii* detritus. Left-hand vertical axis, release of particulate ^{14}C due to pronase E addition to uniformly ^{14}C -labelled detritus (open circles, control; filled circles, pronase E addition). Right-hand vertical axis, dissolution of particulate silica caused by pronase E to unlabelled detritus (open squares, control; filled squares, pronase E addition). Control incubations consisted of diatom detritus suspended in 0.2- μm -filtered, autoclaved sea water (no pronase E added).

state appear to be important variables in bacteria-mediated silicon regeneration, as assemblages collected on different days yielded ~ 4 -fold variation in initial regeneration rate (Table 1). Diatom aggregation into marine snow may also enhance silica solubilization²⁴, probably because aggregates harbour high bacterial concentrations expressing intense ectohydrolase activities⁷. Finally, trophic interactions, such as protozoan grazers and viruses controlling bacteria abundance and activity, may regulate silicon regeneration. Thus, bacteria mediate potentially rapid but highly variable silicon regeneration rates which are regulated not only by temperature but also by multifaceted *in situ* controls on bacterial action on diatom frustules.

Bacteria-mediated silicon regeneration rates from our model system can not *a priori* be extrapolated to the ocean, yet our results provide mechanistic insights on control of diatom productivity as well as illustrating the potential for a greater importance of diatoms in oceanic carbon biogeochemistry. High and variable D:P ratios (0.5–0.8) for open-ocean systems^{5,11} can now be interpreted as being caused mainly by bacteria (because we found little abiotic or protozoa-mediated dissolution, and because metazoa did not cause dissolution¹⁵). Protozoa grazing on nanoplanktonic diatoms may contribute but this is unstudied). Our estimates of silicon regeneration (40–73%) together with field D:P ratios indicate that bacteria-mediated silicon regeneration is of quantitative significance for oceanic diatom biogeochemistry. It has implications for food-web structure and carbon cycling, particularly in open-ocean upwelling high-nutrient low-chlorophyll (HNLC) systems like the equatorial upwelling zone. Upwelled silicic acid is predicted to set the upper limit on diatom production by a model which assumes no silicon regeneration ('silicate pump'²⁴). A D:P ratio of 0.5–0.8 (that is, 50–80% diatom production based on regenerated silicon) would reduce the diatom *f*-ratio (new production/total production) from 1.0 (predicted by the model) to 0.2–0.5 and would increase the predicted upper limit on diatom carbon fixation by 2–5 times through regenerated production. Depending on the relative supply rates of other nutrients (NO_3^- and Fe), diatom production may instead become N- or Fe-limited. In Monterey Bay and in coastal upwelling systems off Peru, even modest silicon regeneration (for example, a D:P ratio of 0.1) can shift diatom production from Si to

N limitation²⁵; similar reasoning might apply to HNLC waters which may oscillate between Si and Fe limitation. Thus the microbial loop may exert critical controls on diatom production and its biogeochemical fate in these oceanic systems. □

Methods

Silica dissolution measurements. Natural seawater samples taken from Scripps Pier were filtered through Nitex mesh (Tetko; 63- or 5- μm pore sizes), polycarbonate membrane filters (Poretics; 1.0- or 0.2- μm pore size), or Whatman filters (GF/F, 0.7- μm pore size and Anodisc, 0.02- μm pore size). Filtered (0.2- μm) autoclaved sea water was also prepared, and served as the control medium. Incubations were done in 300-ml Nalgene bottles. The spectrophotometric silicomolybdate assay described by Parsons *et al.*²⁶ was used to measure the concentration of dissolved silicic acid in incubated samples. Aliquots were sampled and particulate material was pelleted (3,000g; 15 min; 10°C). The supernatant was assayed for silicic acid concentration. Total silica budgets in samples were determined by a hot sodium hydroxide hydrolysis technique²⁷. All incubations and assays were done in sterile polypropylene containers.

Microscopy. Concentrations of diatoms and bacteria were determined by epifluorescence microscopy (Olympus BH-2). For diatoms, samples were filtered onto 0.8- μm , black Nuclepore filters and counted using auto-fluorescence (312 $\times$ magnification). For bacteria, samples were filtered onto 0.2- μm , black Poretics filters after staining with 4,6-diamidino-2-phenyl indole (1 $\mu\text{g ml}^{-1}$; 10 min) and counted at 1,250 $\times$ magnification. Average colonization intensities were obtained by counting the number of attached bacteria on at least 30 individual diatom cells. Attached bacterial concentrations in diatom detritus microenvironment were calculated by dividing average bacteria per diatom by diatom cell volume.

Bacterial production and ectoprotease measurements. Diatom detritus was exposed to 0.7- μm -filtered sea water. Protease activity and bacterial production were measured periodically in the total suspension and <3.0- μm -filtered samples. Cell-specific rates for attached bacteria were obtained on the basis of the difference between total and <3.0- μm fractions. Protease activities were measured using the fluorogenic substrate L-leucine-7-amino-4-methylcoumarin (ref. 28) added at a final concentration of 100 μM ; samples were incubated at the *in situ* temperature (18°C) in the dark. Assays were done in duplicate and one sample was boiled before addition of substrate and served as a control. Fluorescence was measured using a Hoefer TKO-100 fluorometer and was calibrated with 7-amino-4-methylcoumarin. Attached:free-living bacteria protease concentration ratios were calculated as the ratio of enzyme activity in equal volumes of diatom detritus and the bulk phase. Bacterial growth rates were determined by ^3H -leucine incorporation²⁹ during incubations at the *in situ* temperature (18°C) in the dark. Cell-specific growth rate calculations assumed 20 fg C per cell (ref. 30). Assays were done in triplicate.

Pronase E treatment experiment. Release of carbon was measured for diatom detritus made from metabolically uniformly ^{14}C -labelled diatoms suspended in abiotic (0.2- μm -autoclaved) sea water. Parallel incubations with unlabelled diatom detritus were used to measure silicon regeneration. Pronase E (Sigma) was added to a final concentration of 0.1 unit ml^{-1} . Control samples were incubated in abiotic sea water only. Aliquots were removed periodically, centrifuged to sediment the detritus, and ^{14}C activity (labelled incubations) or silicic acid concentrations²⁶ (unlabelled incubations) were measured in the supernatant.

Received 18 June; accepted 23 October 1998.

- Nelson, D. M., DeMaster, D. J., Dunbar, R. B. & Smith, W. O. Jr Cycling of organic carbon and biogenic silica in the Southern Ocean: estimates of water-column and sedimentary fluxes on the Ross Sea continental shelf. *J. Geophys. Res.* **101**, 18519–18532 (1996).
- Longhurst, A. R. & Harrison, W. G. The biological pump: profiles of plankton production and consumption in the upper ocean. *Prog. Oceanogr.* **22**, 47–123 (1989).
- Lewin, J. C. in *Physiology and Biochemistry of Algae* (ed. Lewin, R. E.) 445–455 (Academic, New York, 1962).
- Dugdale, R. C. & Wilkerson, F. P. Silicate regulation of new production in the equatorial Pacific upwelling. *Nature* **391**, 270–273 (1998).
- Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A. & Quéguiner, B. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Glob. Biogeochem. Cycles* **9**, 359–372 (1995).
- Cole, J. J., Findlay, S. & Pace, M. L. Bacterioplankton production in fresh and saltwater ecosystems: a cross-system overview. *Mar. Ecol. Prog. Ser.* **43**, 1–10 (1988).
- Smith, D. C., Simon, M., Alldredge, A. L. & Azam, F. Intense hydrolytic enzyme activity on marine aggregates and implications for rapid particle dissolution. *Nature* **359**, 139–142 (1992).

8. Smith, D. C., Steward, G. F., Long, R. A. & Azam, F. Bacterial utilization of carbon fluxes during a diatom bloom in a mesocosm. *Deep-Sea Res. II* **42**, 75–97 (1995).
9. Kamatani, A. Dissolution rates of silica from diatoms decomposing at various temperatures. *Mar. Biol.* **68**, 91–98 (1982).
10. Lewin, J. C. The dissolution of silica from diatom walls. *Geochim. Geophys. Acta* **21**, 182–198 (1961).
11. Brzezinski, M. A. & Nelson, D. M. The annual silica cycle in the Sargasso Sea near Bermuda. *Deep-Sea Res.* **142**, 1215–1237 (1995).
12. Nelson, D. M. & Goering, J. J. Near-surface silica dissolution in the upwelling region off northwest Africa. *Deep-Sea Res.* **24**, 65–73 (1977).
13. Brzezinski, M. A. & Nelson, D. M. Seasonal changes in the silicon cycle within a Gulf Stream warm-core ring. *Deep-Sea Res.* **36**, 1009–1030 (1989).
14. Cowie, G. L. & Hedges, J. I. Digestion and alteration of the biochemical constituents of a diatom (*Thalassiosira weissflogii*) ingested by a herbivorous zooplankton (*Calanus pacificus*). *Limnol. Oceanogr.* **41**, 581–594 (1996).
15. Tande, K. S. & Slagstad, D. Assimilation efficiency in herbivorous aquatic organisms—the potential of the ratio method using ^{14}C and biogenic silica as markers. *Limnol. Oceanogr.* **30**, 1093–1099 (1985).
16. Biddanda, B. A. & Pomeroy, L. R. Microbial aggregation and degradation of phytoplankton-derived detritus in seawater. I. Microbial succession. *Mar. Ecol. Prog. Ser.* **42**, 79–88 (1988).
17. Brussaard, C. P. D. *et al.* Effects of grazing, sedimentation and phytoplankton cell lysis on the structure of a coastal pelagic food web. *Mar. Ecol. Prog. Ser.* **123**, 259–271 (1995).
18. Berges, J. A. & Falkowski, P. G. Physiological stress and cell death in marine phytoplankton: induction of proteases in response to nitrogen or light limitation. *Limnol. Oceanogr.* **43**, 129–135 (1998).
19. Brussaard, C. P. D., Noordeloos, A. A. M. & Riegman, R. Autolysis kinetics of the marine diatom *Ditylum brightwellii* (Bacillariophyceae) under nitrogen and phosphorus limitation and starvation. *J. Phycol.* **33**, 980–987 (1997).
20. DeLong, E. Archaeal means and extremes. *Science* **280**, 542–543 (1998).
21. Kröger, N., Bergsdorf, C. & Sumper, M. A new calcium binding glycoprotein family constitutes a major diatom cell wall component. *EMBO J.* **13**, 4676–4683 (1994).
22. Smayda, T. J. The suspension and sinking of phytoplankton in the sea. *Oceanogr. Mar. Biol. Annu. Rev.* **8**, 353–414 (1970).
23. Kamatani, A. & Riley, J. P. Rate of dissolution of diatom silica walls in seawater. *Mar. Biol.* **55**, 29–35 (1979).
24. Brzezinski, M. A., Alldredge, A. L. & O'Bryan, L. M. Silica cycling within marine snow. *Limnol. Oceanogr.* **42**, 1706–1713 (1997).
25. Brzezinski, M. A., Phillips, D. R., Chavez, F. P., Friederich, G. E. & Dugdale, R. C. Silica production in the Monterey, California, upwelling system. *Limnol. Oceanogr.* **42**, 1694–1705 (1997).
26. Parsons, T. R., Maita, Y. & Lalli, C. M. *A Manual of Chemical and Biological Methods for Seawater Analysis* (Pergamon, Oxford, 1984).
27. Werner, D. Die Kieselsäure im Stoffwechsel von *Cyclotella cryptica* Reimann, Lewin und Guillard. *Arch. Mikrobiol.* **55**, 278–308 (1966).
28. Hoppe, H. G. Significance of exoenzymatic activities in the ecology of brackish water: measurements by means of methylumbelliferyl-substrates. *Mar. Ecol. Prog. Ser.* **11**, 299–308 (1983).
29. Smith, D. C. & Azam, F. A simple, economical method for measuring bacterial protein synthesis rates in seawater using ^3H -leucine. *Mar. Microb. Food Webs* **6**, 107–114 (1992).
30. Lee, S. & Fuhrman, J. A. Relationships between biovolume and biomass of naturally derived marine bacterioplankton. *Appl. Environ. Microbiol.* **53**, 1298–1303 (1987).

Acknowledgements. We thank D. C. Smith, M. Hildebrand, M. A. Brzezinski, J. T. Hollibaugh and K. A. Bidle for discussions and suggestions. This work was supported by NSF grants to F.A.

Correspondence and requests for materials should be addressed to F.A. (e-mail: fazam@ucsd.edu).

Cool surface waters of the subtropical North Pacific Ocean during the last glacial

Kyung E. Lee & Niall C. Slowey

Department of Oceanography, Texas A&M University, College Station, Texas 77843-3146, USA

One of the most controversial results of the CLIMAP^{1,2} project's reconstruction of past sea surface temperature (SST) is that large areas of the subtropical Pacific Ocean were warmer during the last glacial period than they are today. This finding has important implications because SST patterns at low to subtropical latitudes strongly influence climate, and SST changes are closely linked with climate fluctuations^{3–5}. Until now, a lack of well-preserved, high-resolution marine sediment cores from the region has hindered efforts to confirm these unexpectedly high ice-age SST estimates. Here we use both the oxygen-isotope compositions and species assemblages of planktonic foraminifera in a shallow-water core with high deposition rates near Hawaii to estimate glacial SST of the subtropical North Pacific Ocean. Contrary to the CLIMAP results², our data indicate that the annual average SST in this region was $\sim 2^\circ\text{C}$ cooler during the last glaciation than it is today. These results help to reconcile the marine SST record with inferences drawn from snowline depressions on Hawaii during the last glacial^{3,6}, and should ultimately yield improved estimates

of global climate sensitivity by providing important new constraints on climate model simulations of ice-age cycles.

Located in the central portion of the subtropical North Pacific Ocean, far from confounding influences of oceanic boundary currents and continental landmasses, the Hawaiian Island region is an ideal site at which to investigate the SST 'problem'. Yet attempts to do this using the marine record (including CLIMAP; W. L. Prell, personal communication) have met very limited success. Deep waters in the Pacific are very corrosive to calcium carbonate, so microfossils deposited at most sea-floor sites are poorly preserved and sedimentation rates are low. Dissolution can alter the relative abundance and geochemistry of microfossils while bioturbation of slowly accumulating sediments can mix microfossils of different ages—both processes distort SST records derived from these proxies. At the few sea-floor sites shallow enough to be unaffected by carbonate dissolution, downslope processes confound the records.

Guided by high-resolution seismic imagery, we have identified a shallow-water core that preserves a detailed history of SST variations since the Last Glacial Maximum (LGM). Core PC17 was collected from a moderately sloped portion of the submarine flank of Oahu, Hawaii, located $\sim 6\text{ km}$ off the island's southwestern shore (21.358°N , 158.190°W). SeaMARC II side-scan sonar images indicate that hemipelagic sedimentation occurs at the site of the core and the site is located just down slope from a fossil patch reef which protects it from downslope processes⁷. Annual variability of modern SST there is the same as that in the open ocean⁸ and is broadly representative of the region⁹. We cannot demonstrate that this was also the case during the last glaciation, but there is no reason to suspect otherwise given the evidence available at present. The water depth from which the core was recovered (503 m) is well above the calcite lysocline depth¹⁰. Oxygen-isotope compositions and radiocarbon dates of the mixed-layer-dwelling planktonic foraminifera *Globigerinoides ruber* provide chronological control and indicate that the core's sedimentary record extends from the late Holocene back through the last glacial period (Fig. 1). Sedimentation rates in PC17 during the Holocene ($>10\text{ cm kyr}^{-1}$) and the last glacial period ($3\text{--}10\text{ cm kyr}^{-1}$) are higher than those of typical deep sea sediment cores from the central North Pacific. Enhanced sedimentation may result from local topographic focusing. The regular pattern of the $\delta^{18}\text{O}$ values, the increase in radiocarbon age

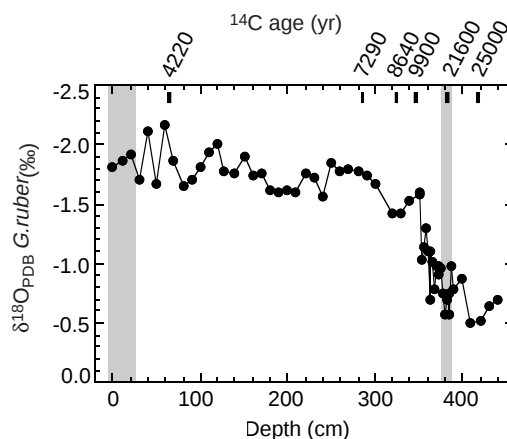


Figure 1 $\delta^{18}\text{O}$ of *Globigerinoides ruber* in core PC17. The white variety of *G. ruber* (212–250 μm) were analysed following standard methods at Texas A&M University and the Woods Hole Oceanographic Institution. Values relative to the PDB standard are plotted versus depth in the core, and represent the average of two or more analyses. Bars at the top of the plot indicate the depth intervals from which *G. ruber* were taken for radiocarbon dating at the National Ocean Sciences AMS Facility (ages were corrected for surface ocean reservoir age by subtracting 400 yr). Shading indicates data from the recent Holocene and the LGM.

with depth, and the absence of sedimentary structures associated with downslope processes are all consistent with continuous sedimentation during both the Holocene and LGM.

The $\delta^{18}\text{O}$ of *G. ruber* reflects the temperature of surface waters near Hawaii. Sediment traps deployed as part of the Hawaii Ocean Time-series Program¹¹ show a sustained period of high flux of *G. ruber* from surface waters during the summer (maximum,

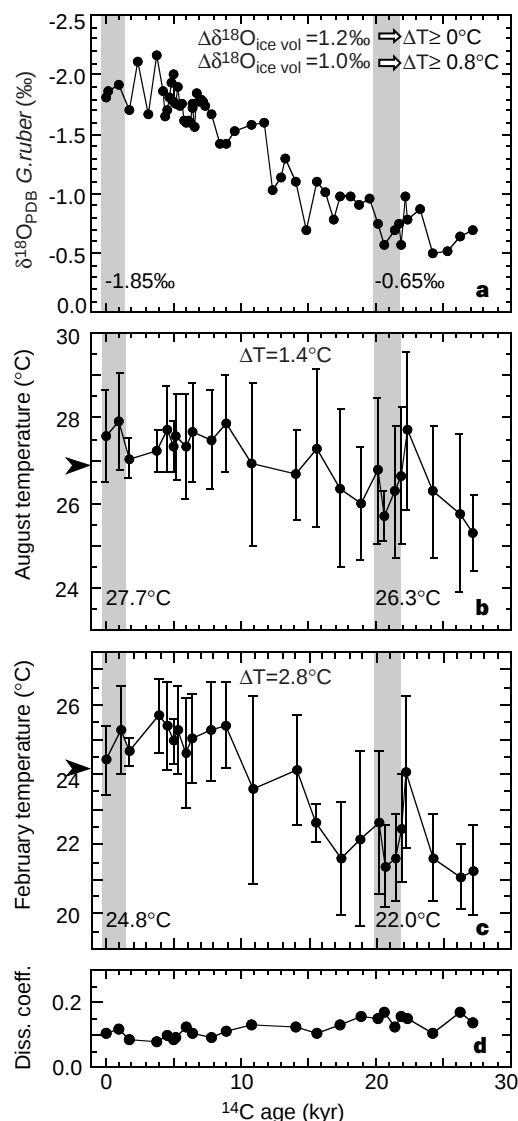


Figure 2 SST proxy records from core PC17. **a**, $\delta^{18}\text{O}$ of *G. ruber* in core PC17 plotted versus radiocarbon age with the average values of recent Holocene and LGM data (shaded, as in Fig. 2). The recent Holocene-LGM difference in SST is estimated assuming that the $\delta^{18}\text{O}$ of mean sea water was 1.0‰ and 1.2‰ greater during the LGM. Changes in the carbonate ion concentration of surface water may allow for the recent Holocene-LGM difference in SST to be up to 1 °C greater than indicated¹⁷. Production of *G. ruber* near Hawaii is greatest during the summer, so these estimates are biased towards summer conditions. **b**, The observed average August SST value (arrow)⁸ and August SST estimated from planktonic foraminifera assemblages using the modern analogue technique (MAT). For each assemblage-based estimate, 300–500 foraminifera (>150 μm) were quantitatively picked and identified following the CLIMAP taxonomy. Vertical lines show the 1 σ standard deviation of temperatures associated with the five closest analogues. **c**, The observed average February SST value (arrow)⁸ and February SST estimated from foraminifera assemblages. **d**, Plot of the dissimilarity coefficient which provides a measure of the similarity between the assemblages in core PC17 and the modern analogues.

~180 shells $\text{m}^{-2} \text{d}^{-1}$) compared to the rest of the year (less than ~45 shells $\text{m}^{-2} \text{d}^{-1}$). We expect the $\delta^{18}\text{O}$ of calcite precipitated in equilibrium¹² with summer surface waters to be about -1.65‰ (based on temperature and $\delta^{18}\text{O}_{\text{water}}$ measured¹³ during summer at nearby GEOSECS Station 206). The mean $\delta^{18}\text{O}$ value of *G. ruber* (white variety, 212–250 μm) observed for the 0–1 kyr interval of core PC17 is ~1.85‰ (Fig. 2a). These values differ by -0.2‰, which agrees with previously reported disequilibrium¹⁴. This suggests that the $\delta^{18}\text{O}$ of Hawaiian *G. ruber* is biased toward summer surface conditions.

If the same seasonal pattern of *G. ruber* production existed during the past, the $\delta^{18}\text{O}$ values of glacial-aged *G. ruber* reflect summer temperatures during the last glaciation. The average $\delta^{18}\text{O}$ of glacial *G. ruber* (-0.65‰) is 1.2‰ greater than the late Holocene value. Taking 1.2‰ as the change in the average isotopic composition of the glacial ocean due to the accumulation of ice¹⁵, then this result indicates that there was no change of summer SST near Hawaii during the last glaciation. Alternatively, if the recent suggestion¹⁶ that the ice volume effect is only 1.0‰ is true, then the summer SST would have been ~1 °C colder. Some uncertainty in these estimates exists because, in addition to the ice-volume effect, changes in regional ocean circulation and climatology could have influenced the late Holocene-glacial difference in the $\delta^{18}\text{O}$ of ocean water. Spero *et al.*¹⁷ have reported that changes in the carbonate-ion concentration of the surface may also influence foraminiferal $\delta^{18}\text{O}$, allowing for glacial SST to be colder by up to another 1 °C.

We also obtained SST estimates by using the modern analogue technique¹⁸ (MAT) to compare assemblages of planktonic foraminifera in core PC17 to a global database of core-top assemblages and average observed SST. The coefficients of dissimilarity (squared chord distance) between the assemblages in core PC17 and the database used to derive estimated SST (Fig. 2d) are small (about 0.1–0.15), indicating that good modern analogues for the Hawaii assemblages exist. Figure 2b and c shows the estimates of August and February SST obtained in this fashion. Average SST estimated for the recent Holocene agrees well with observed Hawaiian values. These results suggest that August SST was ~1.5 °C colder during the last glaciation than today. Within the uncertainties of the MAT and the $\delta^{18}\text{O}$ approaches, this MAT value agrees with the Late Holocene-glacial difference in summer SST estimated from the $\delta^{18}\text{O}$ of *G. ruber*. The MAT results suggest that February SST was nearly 3 °C colder

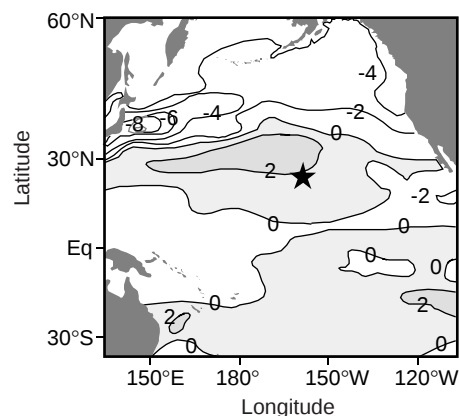


Figure 3 CLIMAP February sea surface temperature (SST) anomalies. The differences between LGM and modern SST estimated² from faunal abundances (shading indicates areas that were warmer during the LGM than today) suggest that a positive anomaly occurs across a large portion of the subtropical North Pacific, with temperatures near the Hawaiian Islands (star) being ~2 °C warmer during the LGM. In contrast, snowline and pollen data from Mauna Kea, Hawaii, suggest that SST was as much as ~4.5 °C cooler if the atmospheric lapse rate did not change^{6,29}.

during the last glaciation than today. Annual average SST was therefore $\sim 2^\circ\text{C}$ colder during the last glaciation.

CLIMAP's global reconstruction² of glacial SST, which was based on assemblages of planktonic foraminifera, indicated that the subtropical North Pacific was warmer during the last glaciation than today (Fig. 3). The same result was obtained when the CLIMAP foraminifera assemblages were re-analysed with the MAT¹⁸. Our results, however, fail to support the existence of anomalously warm SST in the subtropical North Pacific during the last glaciation and instead indicate that these waters were actually colder than today. The magnitude of the cooling is at least as great as marine-sediment-based estimates for cooling of tropical SST in the portion of the equatorial Pacific closest to Hawaii^{2,19}; however, it is not as great as the cooling ($\sim 5^\circ\text{C}$) estimated from geochemical analyses of corals in the western portions of the tropical Pacific²⁰ and Atlantic²¹.

Two independent methods, oxygen-isotope and foraminiferal-assemblage analyses, show that glacial SST in the subtropical North Pacific was on an annual average basis about $1\text{--}2^\circ\text{C}$ colder than today, which is about $3\text{--}4^\circ\text{C}$ colder than suggested by CLIMAP². Why does this difference with CLIMAP² exist? It is not a consequence of the methods employed because CLIMAP and our study used similar faunal-based techniques¹⁸; we suggest that it reflects the sedimentary records considered by the two studies. Samples available to CLIMAP were limited to sediments which accumulated slowly on the sea floor in the presence of corrosive Pacific Deep Water, so effects of bioturbation and dissolution may well have biased their results. These processes are not likely to have biased the sedimentary record in core PC17. It possesses a much higher sedimentation rate and was recovered from a shallow sea-floor site where ambient sea water is saturated with respect to calcite²². For these reasons, the results presented here are more assured than the earlier marine-based estimates of glacial SST in the subtropical North Pacific.

Observational evidence for the distribution of glacial SST provides an important standard for evaluating the sensitivity of general circulation models^{23,24}. Our results, which allow for lower annual average SST in subtropical latitudes during the last glaciation, need to be considered when attempting to reconcile glacial SST values inferred from palaeoclimatic evidence and predicted by models. The distribution of surface temperatures is also tightly linked to the heat transport and general circulation of the oceans. Model simulations of the last glaciation with CLIMAP SST predict that northward heat transport in the Pacific was twice the modern value^{4,25}. When glacial heat transport is prescribed to be similar to modern values, the SST in low to subtropical latitudes is $5\text{--}6^\circ\text{C}$ colder than indicated by CLIMAP⁴. As our data indicate that subtropical SST was about $3\text{--}4^\circ\text{C}$ colder than previously thought², Pacific heat transport during the last glaciation may actually have been only slightly greater than it is today, and the general pattern of intermediate- and deep-ocean circulation need not have differed significantly from what it is today.

That glacial SST in the subtropical North Pacific was colder than today has implications for our understanding of glacial climate. Both today and during the last glaciation, changes in North Pacific SST affect the climate and atmospheric circulation in the extra-tropical Northern Hemisphere^{3,5,26–28}. Of particular relevance to our results, model simulations by Peteet *et al.*⁵ indicate that a 2°C cooling of North Pacific SST would result in significant cooling of air temperatures over North America as well as parts of Europe and Asia. Colder subtropical Pacific surface waters are less of an evaporative source of water vapour to the global atmosphere, leading to a decrease in the atmospheric greenhouse capacity and contributing to global cooling^{4,5}.

The largest discrepancy between glacial SST estimated by CLIMAP and other studies occurs in the subtropical Pacific. While CLIMAP² suggested that SST was nearly 2°C warmer during the last glaciation than today (Fig. 3), the depression of mountain snowlines on Mauna Kea during the LGM allows (after

accounting for lower glacial sea level) for SST near Hawaii to have been nearly 4.5°C colder than today^{6,29}. When Rind and Peteet³ prescribed glacial SST in a global circulation model to be 2°C lower than suggested by CLIMAP², the elevation of the freezing level, and hence the snowline, was depressed by $\sim 500\text{ m}$ at Hawaii. Given that our results indicate that during the LGM the average annual SST at Hawaii was even $\sim 2^\circ\text{C}$ colder than prescribed in the model, it is conceivable that snowlines would have been depressed a further few hundred metres—to, or near to, the elevation inferred from physical evidence on Mauna Kea^{6,29}. Any change in local lapse rate required to completely reconcile our glacial SST and the snowline elevations would be modest. Though this comparison appears favourable, it is nevertheless difficult to assess the direct effect of local SST variations on the Mauna Kea snowline elevation. The local factors which control this elevation may be forced by elements of the climate, such as the strengths and positions of the trade-wind inversion and the Aleutian low-pressure centre, which originate outside the Hawaiian region^{6,28,29}. □

Received 28 July 1998; accepted 6 January 1999.

1. CLIMAP Project Members. The surface of the ice-age earth. *Science* **191**, 1131–1137 (1976).
2. CLIMAP Project Members. *Seasonal Reconstruction of the Earth's Surface at the Last Glacial Maximum* (Map and Chart Series MC-36, Geol. Soc. Am., Boulder, CO, 1981).
3. Rind, D. & Peteet, D. Terrestrial conditions at the Last Glacial Maximum and CLIMAP sea-surface temperature estimates: are they consistent? *Quat. Res.* **24**, 1–22 (1985).
4. Webb, R., Rind, D. H., Lehman, S. R., Healy, R. J. & Sigman, D. Influence of ocean heat transport on the climate of the Last Glacial Maximum. *Nature* **385**, 695–699 (1997).
5. Peteet, D., Del Genio, A. & Lo, K. K.-W. Sensitivity of northern hemisphere air temperature and snow expansion to North Pacific sea surface temperatures in the Goddard Institute for Space Studies general circulation model. *J. Geophys. Res.* **102**, 23781–23791 (1997).
6. Porter, S. C. Hawaiian glacial ages. *Quat. Res.* **12**, 161–187 (1979).
7. Tsutsui, B., Campbell, J. F. & Coulborn, W. T. Storm-generated, episodic sediment movements off Kahe Point, Oahu, Hawaii. *Mar. Geol.* **76**, 281–299 (1987).
8. Hawaii Ocean Time-series database (cited 28 June 1998) (<http://mana.soest.hawaii.edu/pub/hot/ctd/>).
9. Woodruff, S. D., Slutz, R. J., Jenne, R. L. & Steurer, P. M. A comprehensive ocean-atmosphere data set. *Bull. Am. Meteorol. Soc.* **68**, 1239–1250 (1987).
10. Broecker, W. S. & Takahashi, T. The relationship between lysocline depth and in situ carbonate ion concentration. *Deep-Sea Res.* **25**, 65–95 (1978).
11. Dulaney, T. S. & Sautter, L. R. Sedimentation of planktonic foraminifera: seasonal changes in shell flux north of Oahu Hawaii. *Eos* **76**, 85–86 (1996).
12. Erez, J. & Lutz, B. Experimental paleotemperature equation for planktonic foraminifera. *Geochim. Cosmochim. Acta* **47**, 1025–1031 (1983).
13. GEOSECS Executive Committee. *GEOSECS Atlantic, Pacific, and Indian Ocean Expeditions Vol. 7, Shorebased Data and Graphics* (National Science Foundation, Washington DC, 1977).
14. Deuser, W. G. Seasonal and interannual variations in deep-water particle fluxes in the Sargasso Sea and their relation to surface hydrography. *Deep-Sea Res.* **33**, 225–246 (1986).
15. Fairbanks, R. G. A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. *Nature* **342**, 637–642 (1989).
16. Schrag, D. P., Hampt, G. & Murray, D. W. Pore fluid constraints on the temperature and oxygen isotopic composition of the glacial ocean. *Science* **272**, 1930–1932 (1996).
17. Spero, H. J., Bijma, J., Lea, D. W. & Bemis, B. E. Effect of seawater carbonate concentration on foraminiferal carbon and oxygen isotopes. *Nature* **390**, 497–500 (1997).
18. Prell, W. L. *The Stability of Low-latitude Sea-surface Temperatures: an Evaluation of the CLIMAP Reconstruction with Emphasis on the Positive SST Anomalies* (Rep. No. TR025 (DOE/ER/60167-1), US Dept of Energy, Washington DC, 1985).
19. Lyle, M., Pahl, F. & Sparrow, M. Upwelling and productivity changes inferred from a temperature record in the central equatorial Pacific. *Nature* **355**, 812–815 (1992).
20. Beck, J. W., Récy, J., Taylor, F. & Edwards, R. E. Abrupt changes in early Holocene tropical sea surface temperatures derived from coral records. *Nature* **385**, 705–707 (1997).
21. Guilderson, T. P., Fairbanks, R. G. & Rubenstein, J. L. Tropical temperature variations since 20,000 years ago: modulating interhemispheric climate change. *Science* **263**, 663–665 (1994).
22. Sabine, C. L. & Mackenzie, F. T. Bank-derived carbonate sediment transport and dissolution in the Hawaiian Archipelago. *Aquat. Geochem.* **1**, 189–230 (1995).
23. Manabe, S. & Broccoli, A. J. A comparison of climate model sensitivity with data from the last glacial maximum. *J. Atmos. Sci.* **42**, 2643–2651 (1985).
24. Ganopolski, A., Rahmstorf, A., Petoukhov, V. & Claussen, M. Simulation of modern and glacial climates with a coupled global model of intermediate complexity. *Nature* **391**, 351–356 (1998).
25. Miller, J. R. & Russell, G. L. Ocean heat transport during the last glacial maximum. *Paleoceanography* **4**, 141–155 (1989).
26. Horel, J. D. & Wallace, J. M. Planetary-scale atmospheric phenomena associated with the southern oscillation. *Mon. Weath. Rev.* **109**, 813–829 (1981).
27. Wallace, J. M. & Gutzler, D. S. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Weath. Rev.* **109**, 784–812 (1981).
28. COHMAP Project Members. Climatic changes of the last 18,000 years: observations and model simulations. *Science* **224**, 1043–1052 (1988).
29. Porter, S. C. Late Pleistocene eolian sediments related to pyroclastic eruptions of Mauna Kea volcano, Hawaii. *Quat. Res.* **47**, 261–276 (1997).

Acknowledgements. We thank C. Sherman, B. Tsutsui and J. Tribble for help with sampling and discussions; M. Yeager for laboratory assistance; P. Howell, A. Martin and W. Prell for help with foraminifera taxonomy and MAT software; the National Ocean Sciences AMS Facility at Woods Hole for radiocarbon analyses; J. F. Campbell, C. Charles, T. Crowley, L. Keigwin and L. Sautter for discussions; and R. Webb for comments and suggestions. This work and core curation were supported by the NSF.

Correspondence and requests for materials should be addressed to N.C.S. (e-mail: slowey@ocean.tamu.edu).

Holocene periodicity in North Atlantic climate and deep-ocean flow south of Iceland

Giancarlo G. Bianchi & I. Nicholas McCave

Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK

Climate fluctuations during the past millennium are relatively well documented¹. On a longer timescale, there is growing evidence of millennial-scale variability of Holocene climate, at periodicities of ~2,500 and 950 years (possibly caused by changes in solar flux)^{2,3} and ~1,500 years (maybe related to an internal oscillation of the climate system)^{4–6}. But the involvement of deep water masses in these Holocene climate changes has yet to be established. Here we use sediment grain-size data from the Iceland basin to reconstruct past changes in the speed of deep-water flow. The study site is under the influence of Iceland–Scotland Overflow Water (ISOW), the flow of which is an important component of the ‘thermohaline’ circulation that modulates European climate. Flow changes coincide with some known climate events (the Little Ice Age and the Mediaeval Warm Period), and extend over the entire Holocene epoch with a quasi-periodicity of ~1,500 years. The grain-size data indicate a faster ISOW flow when the climate of northern Europe is warmer. However, a second mode of operation is observed in the early Holocene, when warm climate intervals are associated with slower ISOW flow. At that time the melting remnant of land-based, glacial-age ice may have provided a sufficient source of fresh water to the ocean to reduce ISOW flow south of Iceland.

The northeast Atlantic Ocean is of key importance for modulating climate on glacial–interglacial timescales because of heat loss to the atmosphere from the North Atlantic Current and its continuation through the Norwegian Sea, and the convective formation of deep water in the Nordic/Arctic seas, which together provide a temperate climate to northwestern Europe. The poleward flux of warm saline Atlantic waters is partly counterbalanced and maintained by the deep, dense, return flow of ISOW which crosses the Iceland–Scotland ridge mainly through the Faeroe Bank channel to enter the Iceland basin. Together with a similar dense overflow through the Denmark Strait, this process constitutes the initial step of the global ocean conveyor-belt model⁷. The thermohaline conveyor is unstable, and is sensitive to the fresh water/salt balance of the region⁸. Marine isotope stage 3 (~60–30 kyr before present, BP) was characterized by extreme millennial-scale climate instability, resulting in the Dansgaard–Oeschger temperature fluctuations seen in ice-core and marine records of temperature, and also resulting in ice rafting⁹. Our results show that even during the generally stable Holocene there is an underlying fluctuation in the strength of ISOW flow south of Iceland with a similar periodicity, which may be linked to climate changes.

The data we present have been obtained from kasten core NEAP-15K (56°21.92' N, 27°48.68' W) recovered from 2,848 m depth near the crest of Gardar drift in the south Iceland basin (see map in Fig. 1 inset). In this area Holocene sedimentation rates are greatly enhanced (Fig. 1b) owing to the interaction of bottom currents with the sea-bed topography, and the core contains a ~450-cm-long Holocene sediment record. We use a sedimentological near-bottom palaeocurrent speed proxy, the ‘sortable silt’ mean size¹⁰ (SS is the mean grain size of the 10–63-μm terrigenous silt fraction, a parameter that varies independently of sediment supply in current-sorted and deposited muds and for which higher values represent relatively greater near-bottom flow speeds). The mean

Holocene deposition rate for the northeast Atlantic is ~4 cm kyr⁻¹ (ref. 12). The core site is far from any sources of downslope mass transport, but is under the deep western boundary current of the Iceland basin which carries material from south Iceland and its continental slope¹³ to create the ~1,100-km-long Gardar drift¹⁴. The very high sedimentation rates and particle sizes are thus primarily current-controlled, although fine sediment supply by ice rafting during the Holocene cannot be ruled out altogether. However, its effects would be negligible in core NEAP-15K as its Holocene accumulation rates for the terrigenous fine fraction (<63 μm) average ~20 g cm⁻² kyr⁻¹ with peaks of ~80 g cm⁻² kyr⁻¹, whereas in full glacial conditions within the main ice-rafting belt just south of our site (at 51°07.0' N, 21°52.0' W) the flux from the sea surface for the same sediment component was only 1.5 g cm⁻² kyr⁻¹ (ref. 15). The accelerator mass spectrometry (AMS) ¹⁴C dates used for the age model in Fig. 1c have been converted to calendar years BP¹⁶ (where ‘present’ is AD 1950) after applying a 400-year reservoir correction. High-resolution dating was also successfully employed to splice the topmost part of NEAP-15K, which overpenetrated by 35 cm and scrambled the upper sediment section, with box core NEAP-16B recovered from the same site, thereby extending the record to the present day. Sedimentation rates (Fig. 1b) are fairly

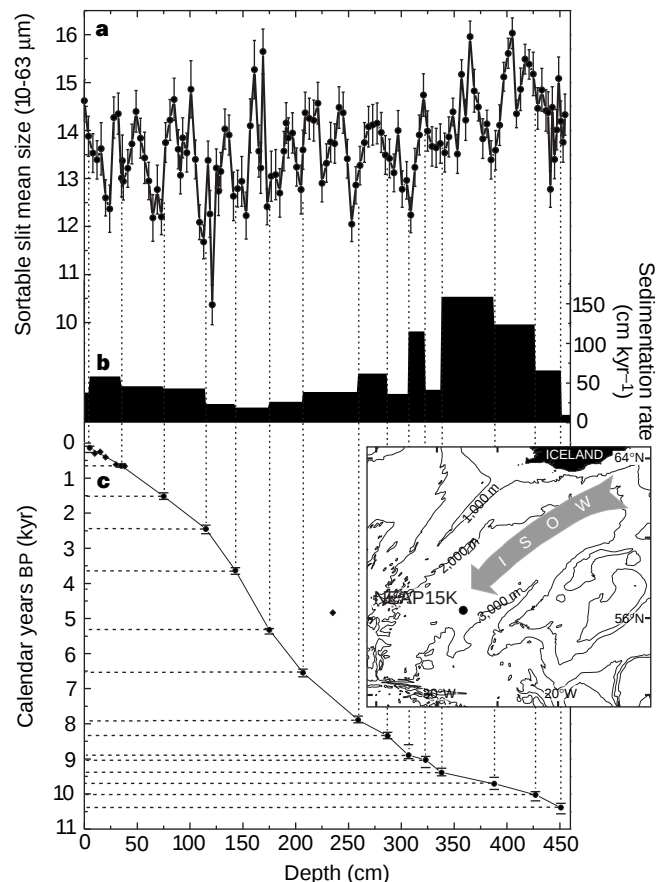


Figure 1 Data from kasten core NEAP-15K with an inset map showing its location in the south Iceland basin and the simplified regional flow of ISOW. **a**, Sortable silt mean size with error bars after ref. 18. **b**, Sedimentation rates between calendar years coinciding with ¹⁴C AMS data used in age model. **c**, Filled circles represent ¹⁴C AMS dates used to construct the age model, and filled diamonds are dates not included in the age model which were predominantly used to splice the uppermost part of NEAP-15K with the bottom of box core NEAP-16B. All dates are from monospecific assemblages of the planktonic foraminifera species *Globigerina bulloides* in the >150-μm fraction. The error range for the calendar dates is represented by horizontal bars above and below the solid circles and is defined as 2 standard deviations.

constant from 9.4 kyr ago to the present, but are significantly higher between ~ 10.4 and 9.4 kyr BP ($\sim 70\text{--}160\text{ cm kyr}^{-1}$) due to the mobilization of sediments accumulated in glacial times by renewed ISOW flow¹⁷. The $\overline{SS}$ varies over the range ~ 10.5 to $16\text{ }\mu\text{m}$ (Fig. 2b), implying that sediment-sorting currents of significantly varying magnitude characterized the Holocene history of ISOW flow. We also note that a number of the fluctuations are small ($\sim 2\text{ }\mu\text{m}$ peak–trough but, nevertheless, significantly bigger than the precision of the method¹⁸), suggesting that some of the changes were subtle, probably reflecting the overall stable nature of sedimentation over the past 10,000 years.

The documented history of climate change in northern Europe over the past few millennia is marked by the alternation of cooler and warmer periods. At present we are still recovering from a time of colder climate known as the Little Ice Age¹⁹, centred at $\sim 400\text{ yr BP}$, which, in our record, coincides with reduced ISOW flow intensity (Fig. 2b). The $\overline{SS}$ shows that since then, deep-water flow vigour has been increasing. Modern values are comparable to the last warm interval in European history, known as the Mediaeval Warm Period, which peaked at different times in various regions surrounding the North Atlantic basin between ~ 750 and $1,050\text{ yr BP}$ (AD ~ 900 to 1250)^{1,20}. The climatic history in the few millennia before the Mediaeval Warm Period is less clear, but Europe appears to have enjoyed a warmer spell at $\sim 2,000\text{ yr BP}$, also referred to as the Roman Warm Period, followed by cooling and glacier advance in the Dark Ages (AD ~ 500 to 1000)^{1,2}. In our record, a peak in deep-current speed centred at $1,850\text{ yr BP}$ (AD 100) coincides with the Roman Warm Period. From these observations and the flow speed fluctuations revealed by Fig. 2b we infer that periods comparable to the Little Ice Age and the Mediaeval Warm Period were a recurrent feature of earlier parts of Holocene climatic history, with the warm intervals coinciding with faster near-bottom water flow in the south Iceland basin.

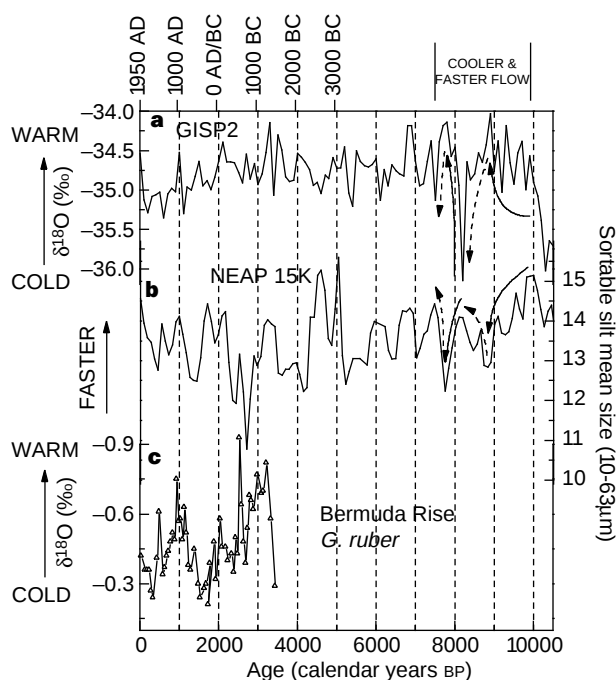


Figure 2 North Atlantic Holocene palaeoenvironmental proxy records on a calendar years BP (and AD/BC) basis. **a**, $\delta^{18}\text{O}$ data from central Greenland GISP2 ice core with gaussian interpolation using a 300-yr window. Solid and dashed arrows between ~ 10 and 7.5 kyr BP represent periods of general warming or cooling which match relative decreases or increases in the flow intensity of ISOW vigour (**b**), respectively. **b**, Sortable silt mean size record for NEAP-15K with gaussian interpolation using a 300-yr window. **c**, Planktonic foraminiferal $\delta^{18}\text{O}$ data from the Sargasso Sea (ref. 26), which mainly reflects changes in sea surface temperature.

The Holocene $\delta^{18}\text{O}$ record in the GISP2 Greenland ice core, which is thought to reflect mainly palaeotemperature changes²¹, reaches typical Holocene values shortly after 10 kyr BP (Fig. 2a). From that time until $\sim 7.5\text{ kyr BP}$ the broad temperature trends of the ice core are similar to our $\overline{SS}$ results (Fig. 2b), with intervals of lower temperatures over Greenland coinciding with times of more vigorous ISOW flow (we note the cycle between 8.9 and 7.5 kyr BP). After $\sim 7.5\text{ kyr BP}$, no clear or consistent relationship exists between the two (note that the faster flow during the cold “ 8.2 kyr event ”²² is well constrained within the 2σ range of dated points in Fig. 1a, c). We suggest that, because sea level was rising at a sustained rate until and just beyond 7.5 kyr BP ²³, the flux of melt water to the northern North Atlantic must have been greater during warmer times; this would have had the effect of reducing the density and hence slowing down the flow of ISOW in the Iceland basin. The system flipped into its present state after $\sim 7.5\text{ kyr BP}$, when most of the remaining glacial ice had been melted and changes in temperature did not cause significant variations in the freshwater budget of the North Atlantic region.

The $\overline{SS}$ record was spectrally analysed (Fig. 3) following gaussian interpolation in the time domain using a 300-yr window and a 90-yr sampling interval (Fig. 2b). Only one broad but pronounced spectral density peak centred at $1,500\text{ yr}$ was obtained. This frequency of fluctuations in the intensity of ISOW flow is very similar to that of ice-rafter events recently reported for the past $\sim 30,000\text{ yr}$ in the northeast Atlantic centred at $1,470\text{ yr}$ (ref. 4), to that of precipitation in western Canada⁵, and to that of monsoon-related aridity/humidity cycles in Arabian dust (which have a $1,450\text{--}1,470\text{-yr}$ period⁶); it is also similar to one of the periodicities recorded in the sea surface and deep-water geochemical and faunal data from the Feni drift between 500 and 340 kyr BP ²⁴; and, finally, it is comparable with the $1,450\text{-yr}$ period present in the GISP2 ice-core chemical data produced by variation in wind strength and storminess due to changing atmospheric circulation patterns over the past 110 kyr (ref. 25). Our results therefore lend strong support to the idea that, as in glacial times⁷, deep-water masses originating in the high-latitude North Atlantic play an important role in modulating climate in the present interglacial.

We also observe several similarities in the fluctuations between the $\overline{SS}$ record and a high-resolution history of surface-water foraminiferal $\delta^{18}\text{O}$ in the Sargasso Sea²⁶ for the past 3.5 kyr (Fig. 2b, c). The correlation, which shows $\delta^{18}\text{O}$ minima (warm) coinciding with periods of faster ISOW flow and vice versa, is $r = 0.47$ for the past $\sim 1,500\text{ yr}$ and 0.76 for the past $\sim 1,200\text{ yr}$ after introducing a lead of 90 yr to the $\overline{SS}$ data (this is acceptable as it is

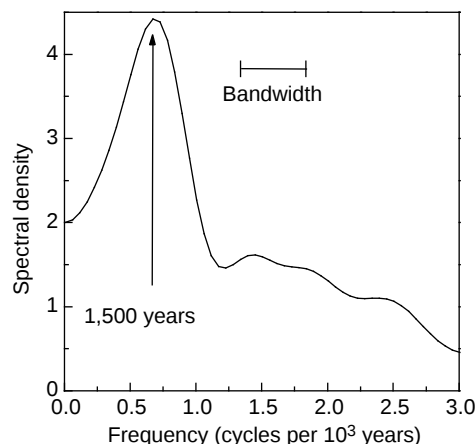


Figure 3 Spectral analysis by the Blackman-Tukey technique²⁴ of the sortable silt mean size record from NEAP-15K using data as shown in Fig. 2b. The $1,500\text{-yr}$ peak accounts for 26% of the total signal in the range above the Nyquist frequency ($1/180\text{ yr}^{-1}$) analysed (including red noise).

still within the errors of the dating techniques employed). Modern oceanographic studies show that, on decadal timescales under the influence of the North Atlantic Oscillation (NAO), convective overturning in the Sargasso and Greenland seas are linked²⁷. However, we cannot explain in the same terms the apparent connection between these two regions on much longer, millennial timescales because the available evidence suggests that varying deep-water convection in the modern Greenland Sea does not affect the intensity of the overflows south of the Greenland–Scotland ridge²⁸. The present 1990s extreme of the NAO has only slight convection in the Greenland–Norwegian seas area, but no diminution in the overflows^{27,29}. This may be because a large component of ISOW volume flux is thought to originate from the North Atlantic Current entering the Arctic via cooling in the Barents Sea, and returning through Fram Strait, without any significant interaction with the Greenland gyre³⁰. Larger-scale atmospheric feedbacks and reorganizations must be invoked to explain the connections observed between the surface northwest Atlantic and the flow intensity of ISOW, particularly in view of the pervasive ~1,500-yr periodicity we observe in our data and which is found in a variety of other proxies from the North Atlantic region^{4,5,25} (in glacial as well as interglacial times) and from as far away as the Arabian Sea⁶.

There is now no doubt that the Earth's climate is highly unstable on millennial timescales. However, our data do not provide an explanation for the ultimate forcing mechanism(s) of these events, and more than one process could be responsible for the changes in ISOW flow vigour recorded in core NEAP-15K. The speed of the overflow south of Iceland could be controlled by the influence of changes in either the salt- or freshwater flux on convection in the Nordic seas; alternatively, a shift in the density of shallow and intermediate water masses entrained by ISOW after overflowing the Iceland–Scotland ridge could also be responsible for the fluctuations of Fig. 2b³¹. Despite the problems in identifying the exact origin of the ISOW flow signal, our data shed light on both the sensitivity and the various operational modes of the thermohaline current system. In an initial early Holocene mode, freshwater flux from the land-based ice melted by warmer conditions is argued to be associated with reduced deep-water activity³², resulting in a less dense and more sluggish ISOW flow over Gardar drift. Then, for the past 7.5 kyr, higher temperatures are in phase with more vigorous deep-water flow, under conditions similar to those at present with little melt water and an oscillation of the supply (by the North Atlantic Current) of saline water and heat to the Nordic seas. So far, no clear 1,500-yr periodicity in either ¹⁴C or ¹⁰Be has been reported from ice cores, suggesting that solar variation is an unlikely forcing mechanism and that an oceanic internal oscillation in 'conveyor' strength is more probable.

The main concern for future climate must be that a possible increase in melting of the Greenland ice sheet resulting from anthropogenically induced atmospheric warming may reach a critical level where the 'conveyor belt' will flip to its early Holocene operational mode³³. The resulting perturbations could conceivably result in climate extremes exceeding those of the Little Ice Age for northern Europe. Without such perturbations, the climate looks likely to be warm for several hundred years (ref. 6). □

Received 29 June; accepted 7 December 1998.

- Lamb, H. H. *Climatic History and the Future* (Princeton Univ. Press, 1985).
- Röthlisberger, F. *10,000 Jahre Gletschergeschichte der Erde* (Sauerländer, Aarau, 1986).
- O'Brien, S. R. et al. Complexity of Holocene climate as reconstructed from a Greenland ice core. *Science* **270**, 1962–1964 (1995).
- Bond, G. et al. A pervasive millennial-scale cycle in North Atlantic Holocene and glacial climates. *Science* **278**, 1257–1266 (1997).
- Campbell, I. D., Campbell, C., Apps, M. J., Rutter, M. W. & Bush, A. B. G. Late Holocene ~1500 year climatic periodicities and their implications. *Geology* **26**, 471–473 (1998).
- Sirocko, F., Garbe-Schonberg, D., McIntyre A. & Molino, B. Teleconnections between the subtropical monsoons and high-latitude climates during the last deglaciation. *Science* **272**, 526–529 (1996).
- Broecker, W. S. & Denton, G. H. The role of the ocean-atmosphere reorganizations in glacial cycles. *Geochim. Cosmochim. Acta* **53**, 2465–2501 (1989).
- Broecker, W. S., Bond, G. & Klas, M. A salt oscillator in the glacial Atlantic? 1. The concept. *Paleoceanography* **5**, 469–477 (1990).

- Rasmussen, T. L., Thomsen, E., van Weering, T. C. E. & Labeyrie, L. Rapid changes in surface and deep water conditions at the Faeroe Margin during the last 58,000 years. *Paleoceanography* **11**, 757–771 (1996).
- McCave, I. N., Manigetti, B. & Robinson, S. G. Sortable silt and fine sediment size/composition slicing: parameters for palaeocurrent speed and palaeoceanography. *Paleoceanography* **10**, 593–610 (1995).
- McCave, I. N., Manigetti, B. & Beveridge, N. A. S. Circulation in the glacial North Atlantic inferred from grain-size measurements. *Nature* **374**, 149–152 (1995).
- Balsam, W. L. & McCoy, F. W. Atlantis sediments: glacial/interglacial comparisons. *Paleoceanography* **2**, 531–542 (1987).
- Shor, A. N. Bottom currents and abyssal sedimentation processes south of Iceland. Thesis, Woods Hole Oceanographic Inst. (1980).
- McCave, I. N. & Tucholke, B. E. in *The Western North Atlantic Region* Vol. M (eds Vogt, P. R. & Tucholke, B. E.) 451–468 (Geol. Soc. Am., Boulder, 1986).
- Manigetti, B. & McCave, I. N. Depositional fluxes, palaeoproductivity, and ice rafting in the NE Atlantic over the past 30 ka. *Paleoceanography* **10**, 579–592 (1995).
- Stuiver, M. & Reimer, P. J. Extended ¹⁴C data-base and revised calib 3.0 C-14 age calibration program. *Radiocarbon* **35**, 215–230 (1993).
- Keigwin, L. D. & Jones, G. A. Glacial-Holocene stratigraphy, chronology, and paleoceanographic observations on some North Atlantic sediment drifts. *Deep-Sea Res.* **36**, 845–867 (1989).
- Bianchi, G. G., Hall, I. N., McCave, I. N. & Joseph, L. Measurement of the sortable silt current speed proxy using the Sedigraph 5100 and Coulter Counter II: Precision and accuracy. *Sedimentology* (in the press).
- Grove, J. M. *The Little Ice Age* (Methuen, London, 1988).
- Grove, J. M. & Switsur, V. R. Glacial geological evidence for the Medieval Warm Period. *Clim. Change* **26**, 143–169 (1994).
- Jouzel, J. et al. Validity of the temperature reconstruction from water isotopes in ice cores. *J. Geophys. Res.* **102**, 26471–26488 (1997).
- Alley, R. B. et al. Holocene climatic instability: A prominent, widespread event 8,200 yr ago. *Geology* **25**, 483–486 (1997).
- Bard, E. et al. Deglacial sea-level record from Tahiti corals and the timing of global meltwater discharge. *Nature* **382**, 241–244 (1996).
- Oppo, D. W., McManus, J. F. & Cullen, J. L. Abrupt climate events 500,000 to 340,000 years ago: evidence from subpolar North Atlantic sediments. *Science* **279**, 1335–1338 (1998).
- Mayewski, P. A. et al. Major features and forcing of high-latitude northern hemisphere atmospheric circulation using a 110,000-year-long glaciochemical series. *J. Geophys. Res.* **102**, 26345–26366 (1997).
- Keigwin, L. D. The Little Ice Age and Medieval Warm Period in the Sargasso Sea. *Science* **274**, 1504–1508 (1996).
- Dickson, R. R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
- Dickson, R. R. & Brown, J. The production of North Atlantic Deep Water: Sources, rates and pathways. *J. Geophys. Res.* **99**, 12319–12341 (1994).
- van Aken, H. M. & Becker, G. Hydrography and through-flow in the north-eastern North Atlantic Ocean: the NANSEN project. *Prog. Oceanogr.* **38**, 297–346 (1996).
- Mauritzen, C. Production of dense overflow waters feeding the North Atlantic across the Greenland–Scotland Ridge. Part 1: Evidence for a revised circulation scheme. *Deep-Sea Res.* **143**, 769–806 (1996).
- Price, J. F. & Baringer, M. O'N. Outflows and deep water production by marginal seas. *Prog. Oceanogr.* **33**, 161–200 (1994).
- Sakai, K. & Peltier, W. R. A multibasin model of the global thermohaline circulation: Paleocceanographic analyses of the origins of ice-age climate variability. *J. Geophys. Res.* **101**, 22535–22562 (1996).
- Broecker, W. S. Thermohaline circulation, the Achilles heel of our climate system: Will man-made CO₂ upset the current balance? *Science* **278**, 1582–1588 (1997).
- OS-3 ARAND SYSTEM: Documentation and Examples Vol. 1 (Computer Center, Oregon State Univ., 1973).

Acknowledgements. We thank S. Crowhurst for help in the spectral analysis of the data presented here, N. Shackleton for comments on an earlier draft, and R. Dickson for hydrographic observations. This work was supported by UK NERC for the North East Atlantic Palaeoceanography and Climate Change project (NEAPACC).

Correspondence and requests for materials should be addressed to G.G.B. (e-mail: ggb20@esc.cam.ac.uk).

Auditory distance perception in rooms

Adelbert W. Bronkhorst & Tammo Houtgast

TNO Human Factors Research Institute, PO Box 23, 3769 ZG Soesterberg, The Netherlands

The perceived distance of a sound source in a room has been shown to depend on the ratio of the energies of direct and reflected sound¹. Although this relationship was verified in later studies^{2–4}, the research has never led to a quantitative model. The advent of techniques for the generation of virtual sound sources^{5,6} has made it possible to study distance perception using controlled, deterministic stimuli. Here we present two experiments that make use of such stimuli and we show that a simple model, based on a modified direct-to-reverberant energy ratio, can accurately predict the results and also provide an explanation for the 'auditory horizon' in distance perception. The modification of the ratio consists of the use of an integration time of 6 milliseconds in the

calculation of the energy of the direct sound. This time constant seems to be important in spatial hearing—the precedence effect is also based on a similar integration window⁷.

A well-known phenomenon in auditory distance perception is our inability to accurately judge distances of sound sources in the free field^{8,9}. This deficit is due to the fact that the most important distance cue in those circumstances—loudness—is confounded with the level of the sound itself. Distances in enclosed environments can be estimated more accurately because one 'absolute' cue is then available: the energy ratio of direct and reverberant sound^{3,4}. Apparently, humans have dwelt long enough in such environments for the auditory system to learn to take advantage of this specific cue. Studies of auditory distance perception have been hampered by the difficulty in controlling and quantifying the actual stimulus presented to the listener's ears. By using techniques developed within 'virtual acoustics', our experiments have overcome these limitations.

In the two experiments (see Methods), listeners were presented with bursts of pink noise, convolved with a binaural impulse response^{10,11} and delivered through headphones. The binaural impulse response was designed to simulate the characteristics of the experimental room and the loudspeaker, while removing the influence of the headphone. Listeners were asked to rate the apparent distance of the virtual sound source, created in this way, on an equal-interval scale. Anchor points were distance zero (for in-the-head localization), the distance of a (silent) loudspeaker, placed in front of the listener, and (in the second experiment) distances of two poles positioned between the listener and the loudspeaker. We carried out the experiments in two rooms with different acoustics. Distance estimates were obtained as a function of source distance, overall stimulus level, the number of reflections included in the impulse response, and the relative level of these reflections.

Before the experiments, we checked the fidelity of the simulation by asking listeners to compare the sound of the loudspeaker (with headphones removed) directly with that of a virtual source, generated at the same distance using 800 reflections. All listeners reported that the virtual source coincided with the loudspeaker. Because the experiments included many conditions in which the binaural impulse response was modified drastically, we deemed it important

to ensure that the listener's adaptation to sound source and acoustic environment was maintained. We therefore repeated the 'fidelity check' several times between blocks of trials.

Analysis of the results revealed three main points (Figs 1, 2). First, overall sound level apparently had only a minor effect on the distance judgements. Second, perceived distance was determined mainly by three variables: virtual source distance, number of reflections and relative level of the reflections. Third, when 27 or more reflections were used, a 'horizon' effect occurred: for small distances perceived distance was almost proportional to source distance, but perceived distance increased only slowly when source distance was more than ~2 m. The latter effect has also been found in earlier studies of distance perception²⁻⁴.

It was clear that the 'true' direct-to-reverberant energy ratio could not be used to predict the results: for all conditions with increased level of reflections, the perceived distance should have equalled or exceeded that obtained for 800 reflections. We first attempted to fit the data by including a factor that depended on the duration of the reverberance (the reflections that were actually presented). Predictions were quite accurate but results for the two rooms could be brought together only by adding a parameter that was not related to those normally used in room acoustics. It was also difficult to envisage a neural mechanism that would be able to detect and combine energy and time measures in the manner suggested by such a model.

We obtained a better solution when duration was not inserted directly but was taken into account by an integration window used for determining the energy of the 'direct' sound. The resulting model is simple, in line with what is known about the temporal resolution of the auditory system¹², and yields excellent predictions of the data without requiring extra parameters for the two rooms. Further support for the model is provided by evidence⁷ that a similar integration window is involved in the (closely related) precedence effect—the dominance of the first part of a stimulus in sound localization. The integration window also provides an explanation for the 'auditory horizon': the energy within the window does not decrease further as a function of distance when the energy of the direct sound is small compared with that of the fraction of reverberant sound that is included. The model can be

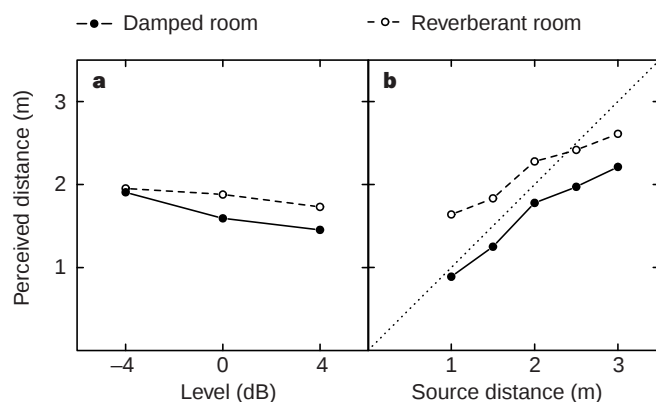


Figure 1 Effects of overall level and source distance on perceived distance (data from the first experiment). **a**, The effect of overall level on perceived distance is relatively small: an 8-dB level increase yields an average decrease in distance of <0.5 m. Data are averaged across all other variables (s.e.m. $\leq$ 0.4 m, $n = 540$). **b**, Effect of source distance on perceived distance. Results for virtual sources created using 800 reflections. In the damped room, sources are consistently judged to be closer than in the reverberant room. The decreasing slope of the curves illustrates the 'auditory horizon' effect. Data are averaged across subjects and presentation levels (s.e.m. $\leq$ 0.12 m, $n = 36$).

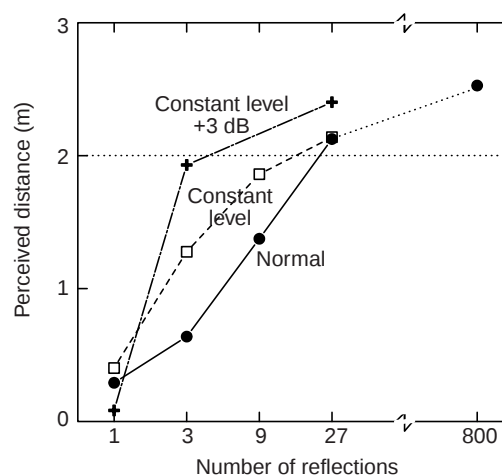


Figure 2 Perceived distance as a function of the number of reflections and their level for a virtual source at 2 m distance in the reverberant room (data from the second experiment). Results are shown for uncorrected reflections (normal), for reflections with a constant level, matched to that of 800 reflections, and for reflections with a constant level raised by an extra 3 dB. Perceived distance increases rapidly as a function of the number of reflections, especially when the reflections are enhanced. Virtual sources created with only one reflection are always judged to be very close. Data are averaged across subjects and presentation levels (s.e.m. $\leq$ 0.12 m, $n = 36$).

expressed as:

$$d_s = Ar_h \left(\frac{\hat{E}_r}{\hat{E}_d} \right)^j \quad (1)$$

where d_s is the subjective distance (in m), A and j are constants, and $\hat{E}_d$ and $\hat{E}_r$ are the (modified) energies of direct and reverberant sounds, respectively. The reverberation radius r_h (the distance at which the energy of the complete reverberance equals that of the direct sound) depends only on the properties of the room and the sound source and can be calculated using¹³:

$$r_h = 0.1 \left(\frac{GV}{\pi T} \right)^{\frac{1}{2}}$$

where V and T are the volume (in m³) and reverberation time (in s) of the room, respectively, and G is the directivity factor of the sound source. The integration window W , used to calculate the energy of the direct sound, contains two parameters, the duration t_w and the slope s . It is given by:

$$W = \begin{cases} 1 & \text{for } 0 < t \leq t_w - \frac{\pi}{2s} \\ 0.5 - 0.5 \sin[s(t - t_w)] & \text{for } t_w - \frac{\pi}{2s} < t < t_w + \frac{\pi}{2s} \\ 0 & \text{for } t \geq t_w + \frac{\pi}{2s} \end{cases}$$

where $t = 0$ is the time of arrival of the direct sound. The energy of the reverberant sound is calculated using the function $1 - W$.

In applying the model to the results of the two experiments (162 distance estimates, obtained after averaging across listeners and presentation levels), four parameters were fitted, namely A , j , t_w and s . We calculated the modified direct-to-reverberant energy ratio for each individual data point using the corresponding impulse response. An iterative least-squares fit yielded a correlation coefficient of 0.94 and the following estimates and standard errors: $A = 1.9 \pm 0.1$; $j = 0.49 \pm 0.03$; $t_w = 6.1 \pm 0.3$ ms; and $s = 4.0 \times 10^2 \pm 0.4 \times 10^2$ s⁻¹. The modified direct-to-reverberant energy ratio, used in the model, is larger than the true ratio because some of the reverberant energy is included in the calculation of the

direct energy. This means that A will generally have a value larger than 1. A will, however, also depend on the degree of adaptation of the listener to the environment and on the stimulus level (the latter influence is difficult to quantify because loudness is a relative distance cue).

We have applied the model to results obtained in three previous studies^{3,4,14}. Because only global parameters are given in these studies, we rewrote equation (1):

$$d_s = Ar_h \left(\frac{1}{r_h^2(1 + Gkt_w)/d^2 + Gkt_w} \right)^{\frac{1}{2}} \quad (2)$$

Here, d_s , A , r_h , G , and t_w are as defined above, $k = 6\ln(10)/T$ and d is the distance of the sound source. In deriving the expression, we used a step function for W and assumed that the energy of the reverberant sound remains constant during a short period (when early reflections arrive from surfaces close to the listener) before decaying exponentially (with time constant $1/k$). Because the early reflections are emitted in the direction of the listener, we multiplied the energy of the constant part by G . Inspection of several calculated impulse responses showed that the period with constant energy could be taken to equal t_w . Unfortunately, G was reported in none of the studies; it was assumed to be equal to 3 in all cases. We fitted the parameter A to the data; its value ranged from 1.2 to 2.7. As shown in Fig. 3, there is a good agreement between the data and the predictions of our model.

The model suggests that listeners combine two sources of information when estimating the distance of a sound source in a room, namely direct perceptual information, which is based on comparison of energies of direct and reverberant sound, and indirect information, which accumulates when the listener is adapting to the sound source and the environment. Because the latter factor, represented in the model by the term Ar_h , depends on a rather complex combination of acoustic parameters (reverberation time, room volume and source directivity), it will be interesting to study this process in more detail: how quick and accurate is it and which information is actually used? Finally we mention a practical application of the model. When creating a realistic virtual auditory environment, simulation of proper distance cues is important but calculation of all reflections presents severe computational difficulties. Our model can be used to determine the minimum number of reflections, and their energies, that should be included. □

Methods

The binaural impulse response used to create a virtual sound source had a maximum duration of 0.11 s. It was calculated by adding short impulse responses (each with a duration of 3.4 ms) for the direct sound and for up to 800 reflections. We used a mirror-images model, which took into account all relevant geometric and acoustic parameters, to determine the properties (delay, angle of incidence and frequency spectrum) of the reflections. The impulse responses were derived from head-related transfer functions, measured in an anechoic room for each listener individually using miniature microphones placed in blocked ear canals¹⁵. Measurements were performed for about 10³ angles of incidence, spanning almost the complete sphere around the listener.

The two rooms in which the experiments were carried out had reverberation times of 0.1 s and 0.5 s, and volumes of 37 m³ and 65 m³, respectively ('damped' room and 'reverberant' room, respectively). Stimuli with a duration of 1 s were generated from virtual sources at distances of 1–3 m. To minimize the confounding effect of level on distance, the overall level was first adjusted such that the energy of the direct sound was independent of distance, and subsequently shifted by -4, 0 or +4 dB. The reference level was ~65 dBA. The number of reflections that was included in the impulse response was limited: only the n strongest reflections were used, with $n = 1, 3, 9, 27, 81$ or 800. The relative level of the reflections was either left unchanged or (for 27 or fewer reflections) increased such that the total level was equal to that of all 800 reflections. In the second experiment, conditions were also included in which the reflections had a relative level 3 dB above that of all 800 reflections.

In the first experiment, the loudspeaker that acted as reference point was

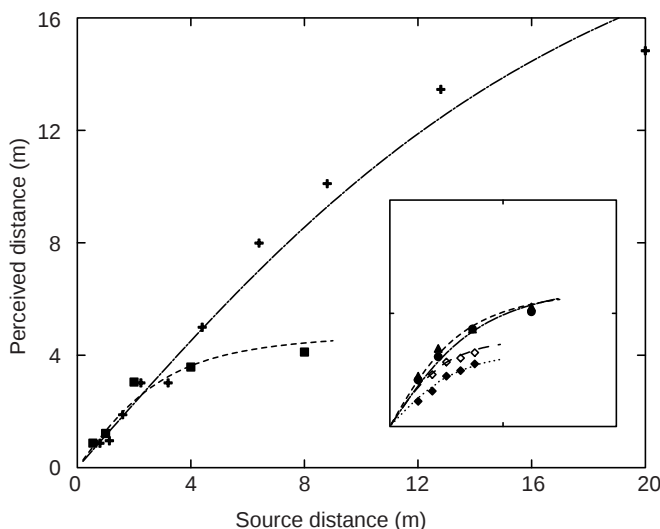


Figure 3 Subjective distances of sound sources reported here and in three previous studies^{3,4,14}, and predictions calculated using equation (2). Predictions are indicated by lines and data points by symbols (squares, ref. 3; plus symbols, ref. 14; triangles, ref. 4, IEC room; circles, ref. 4, classroom; filled diamonds, this study, damped room; open diamonds, this study, reverberant room). Given the wide range of acoustic environments addressed in these studies (T ranges from 0.1 s to 2 s, and room volume from 37 m³ to 9,600 m³), the agreement between the experimental data and our predictions is remarkably good.

placed at a distance of 2 m and listeners could use four intervals (centred at 0, 1, 2 and 3 m) for their distance judgements. Because the low number of intervals could have affected the results (owing to the limited resolution and the possible occurrence of end effects), the interval number was doubled in the second experiment (intervals were now centred at 0–3.5 m in 0.5-m steps). In addition, the loudspeaker was moved to a distance of 3 m to check whether its position had introduced a bias. The 3-m gap was bridged by two poles, placed at distances of 1 and 2 m, that provided additional reference points. Six listeners participated in each experiment. A between-groups analysis of variance, applied to results of the two experiments, showed that the results for the conditions that were replicated did not differ significantly ($P = 0.73$). Thus, results were affected neither by the number of intervals nor by the position of the loudspeaker.

Received 3 December; accepted 21 December 1998.

1. Békésy, G. v. Über die Entstehung der Entfernungsempfindung beim Hören. *Akustische Z.* **3**, 21–31 (1938).
2. Mershon, D. H. & King, L. E. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Percept. Psychophys.* **18**, 409–415 (1975).
3. Mershon, D. H. & Bowers, J. N. Absolute and relative cues for the auditory perception of egocentric distance. *Perception* **8**, 311–322 (1979).
4. Nielsen, S. H. Auditory distance perception in different rooms. *J. Audio Eng. Soc.* **41**, 755–770 (1993).
5. Wightman, F. L. & Kistler, D. J. Headphone simulation of free-field listening: II. Psychophysical validation. *J. Acoust. Soc. Am.* **85**, 868–878 (1989).
6. Bronkhorst, A. W. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.* **98**, 2542–2553 (1995).
7. Houtgast, T. & Aoki, S. Stimulus-onset dominance in the perception of binaural information. *Hearing Res.* **72**, 29–36 (1994).
8. Coleman, P. D. Failure to localize the source distance of an unfamiliar sound. *J. Acoust. Soc. Am.* **34**, 345–346 (1962).
9. Coleman, P. D. An analysis of cues to auditory distance perception in free space. *Psychol. Bull.* **60**, 302–315 (1963).
10. Blauert, J. *Spatial Hearing* (MIT Press, Cambridge, Massachusetts, 1997).
11. Begault, D. R. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *J. Audio Eng. Soc.* **40**, 895–904 (1992).
12. Moore, B. C. J., Glasberg, B. R., Plack, C. J. & Biswas, A. K. The shape of the ear's temporal window. *J. Acoust. Soc. Am.* **83**, 1102–1116 (1988).
13. Kutttruff, H. *Room Acoustics* (Applied Science, London, 1973).
14. Wettchurck, R., Plenge, G. & Lehringer, F. Entfernungswahrnehmung beim natürlichen Hören sowie bei kopfbezogenen Stereophonie. *Acustica* **29**, 260–272 (1973).
15. Møller, H., Sørensen, M. F., Hammershoi, D. & Jensen, C. B. Head-related transfer functions of human subjects. *J. Audio Eng. Soc.* **43**, 300–321 (1995).

Acknowledgements. Part of this research was done within the European AUDIS project.

Correspondence and requests for materials should be addressed to A.W.B. (e-mail: bronkhorst@tm.tno.nl)

The arrangement of the three cone classes in the living human eye

Austin Roorda*† & David R. Williams*

*Center for Visual Science, University of Rochester, Rochester, New York 14627, USA

Human colour vision depends on three classes of receptor, the short- (S), medium- (M), and long- (L) wavelength-sensitive cones. These cone classes are interleaved in a single mosaic so that, at each point in the retina, only a single class of cone samples the retinal image. As a consequence, observers with normal trichromatic colour vision are necessarily colour blind on a local spatial scale¹. The limits this places on vision depend on the relative numbers and arrangement of cones. Although the topography of human S cones is known^{2,3}, the human L- and M-cone submosaics have resisted analysis. Adaptive optics, a technique used to overcome blur in ground-based telescopes⁴, can also overcome blur in the eye, allowing the sharpest images ever taken of the living retina⁵. Here we combine adaptive optics and retinal densitometry⁶ to obtain what are, to our knowledge, the first images of the arrangement of S, M and L cones in the living human eye. The proportion of L to M cones is strikingly different in two male subjects, each of whom has normal colour vision. The

mosaics of both subjects have large patches in which either M or L cones are missing. This arrangement reduces the eye's ability to recover colour variations of high spatial frequency in the environment but may improve the recovery of luminance variations of high spatial frequency.

We measured the eye's aberrations with a Hartmann–Shack wavefront sensor and compensated for them with a deformable mirror (see ref. 4 for details). We then collected images of the cone mosaic, as shown in Fig. 1a, with a charge-coupled device (CCD). Individual cones were classified by comparing images taken when the photopigments were fully bleached with those taken when the photopigments were either dark-adapted or exposed to a light that selectively bleached one photopigment. From these images, we created absorbance images that remove static features to reveal only the distribution of the photolabile pigments that distinguish the cone classes.

To distinguish S cones from M and L cones, we obtained absorbance images from dark-adapted and fully bleached images.

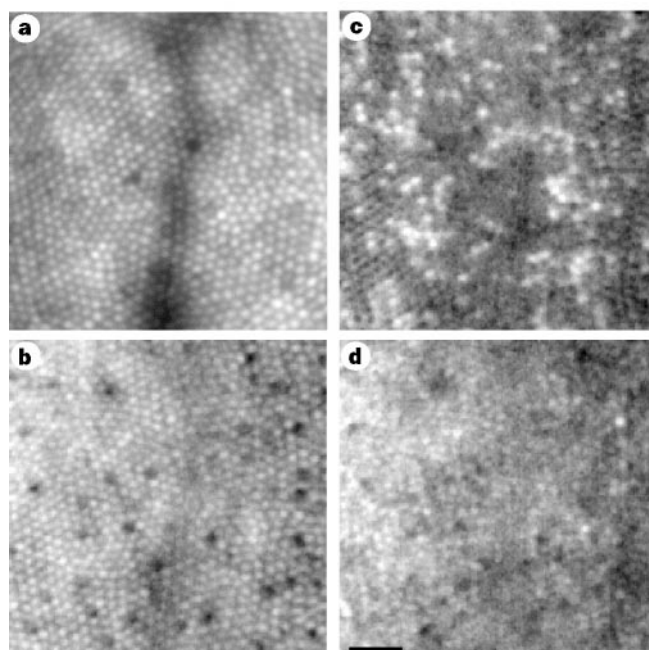


Figure 1 Images of the right eye of subject JW. The mosaic was illuminated with a 4-ms flash (1 degree diameter, $\sim 0.3 \mu\text{J}$) through a 2-mm entrance pupil. Light of wavelength 550 nm was used to maximize absorbance by L- and M-cone photopigment. Images were obtained with a 6-mm exit pupil at a retinal eccentricity of one degree nasal from the foveal centre, which is located to the left of the image. For each retinal location, about 50 images taken over 5 days were averaged to increase the signal-to-noise ratio. Fixational eye movements translate the image from flash to flash, requiring registration with cross-correlation before averaging. **a**, A registered sum of 61 images taken after a full bleach. **b–d**, Absorbance images of the patch of cones shown in **a**, defined as 1 minus the ratio of the absorbance of a dark-adapted or selectively bleached image to the absorbance of the corresponding fully bleached image. Images of fully bleached retina were obtained following exposure to 550-nm light (70-nm bandwidth, 37×10^6 troland-seconds). Images of dark-adapted retina were taken following 5 min of dark adaptation. **b**, The absorbance image of dark-adapted cones, revealing a sparse array of S cones which appear dark because of their low absorbance at 550 nm. **c**, **d**, Absorbance images following a 470-nm (**c**) and 650-nm (**d**) selective bleach. Bleaching energies were set by calculation and then modified empirically to achieve the maximum possible difference between M- and L-photopigment concentration. The absorbance images for both bleaching conditions have higher pigment density toward the fovea; this is caused by the increases in the length of the outer segment and in macular pigment, which reduce bleaching in the 470-nm condition. Scale bar represents 5 arcmin of visual angle.

†Present address: University of Houston, College of Optometry, Houston, Texas 77204-6052, USA

Because S cones absorb negligibly at the imaging wavelength of 550 nm, whereas M and L cones absorb strongly at this wavelength, the S cones appear as a sparse array of dark cones in the absorbance image of Fig. 1b while the M and L cones appear bright. To distinguish L from M cones, we took images immediately after one of two bleaching conditions: dark-adapted retina was exposed to a 650-nm light, which selectively bleached the L pigment, or a 470-nm light, which selectively bleached the M pigment. The absorbance image for the 650-nm bleach reveals dark, low-absorbance L cones that have been heavily bleached and bright, highly absorbing M cones that were spared from bleaching (Fig. 1c). The absorbance image for the 470-nm bleach shows selective bleaching of M cones, but the effect is smaller than the effect of the 650-nm bleach because of the similarity of M-cone and L-cone action spectra at this wavelength (Fig. 1d).

Figure 2a–d shows that, for two trichromats with normal colour vision, the absorbances of single cones after 470-nm and 650-nm bleaches produce a bimodal distribution. If these modes represent L and M cones, then only a single mode should be observed in a similar experiment on a protanope, a person who lacks the L pigment. This prediction is confirmed by the data shown in Fig. 2e, f. Table 1 summarizes the numbers of S, M and L cones found in the two trichromats. The relative number of L and M cones differs greatly between these two subjects. In two patches of retina, one from the nasal and one from the temporal fovea, subject JW had a mean ratio of L to M cones of 3.79 whereas AN had a ratio of 1.15. This large individual difference is consistent with the variability found using psychophysical methods^{7–10}, spectral electroretinograms^{11,12}, microspectrophotometry^{13,14} and messenger RNA analysis^{15,16}.

The arrangements of S, M and L cones for subjects JW and AN are shown in the pseudocolour images in Fig. 3. The distribution of the sparse S cones is not significantly different from random in either trichromat. This agrees with previous results³ that showed that the developmental mechanism used to space S cones in a regular,

Table 1 Cone numbers for two subjects with normal colour vision

Subject	Number of cones	L cones (%)	M cones (%)	S cones (%)	Error (%)	L:M ratio
JW	1462	75.8	20.0	4.2	2.1	3.79
AN	522	50.6	44.2	5.2	5.6	1.15

The error column shows the percentage of cones labelled M (or L) that are actually L (or M) cones. The error in the assignment of L and M cones is taken to be the fractional area of the intersection of the sum of two gaussians used to fit the histograms shown in Fig. 2.

nonrandom manner in humans is not seen near the fovea. Perhaps cone migration during the formation of the fovea disrupts regular S-cone spacing. The assignment of M and L cones is not significantly different from random in JW's eye but in AN's eye the M and L cones are significantly more aggregated than expected in a random mosaic. This additional clumping could arise if, for example, progenitor cells were to bias the 'decisions' of their progeny to express either M or L opsin. However, we cannot exclude the possibility that the departure from a random distribution in AN's eye is a consequence of residual optical blur in his retinal images. Optical blur could increase the chances of misclassifying especially those cones that are surrounded by cones of the opposite class, and this would exaggerate clumping. The overlap in the two distributions of Fig. 2d indicates that about 6% of AN's M and L cones were misidentified. Simulations suggest that this low error rate may nonetheless be high enough to produce the tendency towards the additional clumping we observed. Microspectrophotometry¹⁷ on small foveal patches of excised talapoin retina indicated a random assignment of M and L cones and photopigment transmittance imaging in a macaque peripheral retina¹⁸ showed a tendency towards aggregation of M and L cones. All studies agree that there is no tendency for the M cones to disperse themselves uniformly among the L cones, and the lack of a regular packing scheme for these cone classes may be ubiquitous among old-world primates.

What are the implications of this arrangement for vision? The coarse grain of the cone submosaics causes fluctuations in the

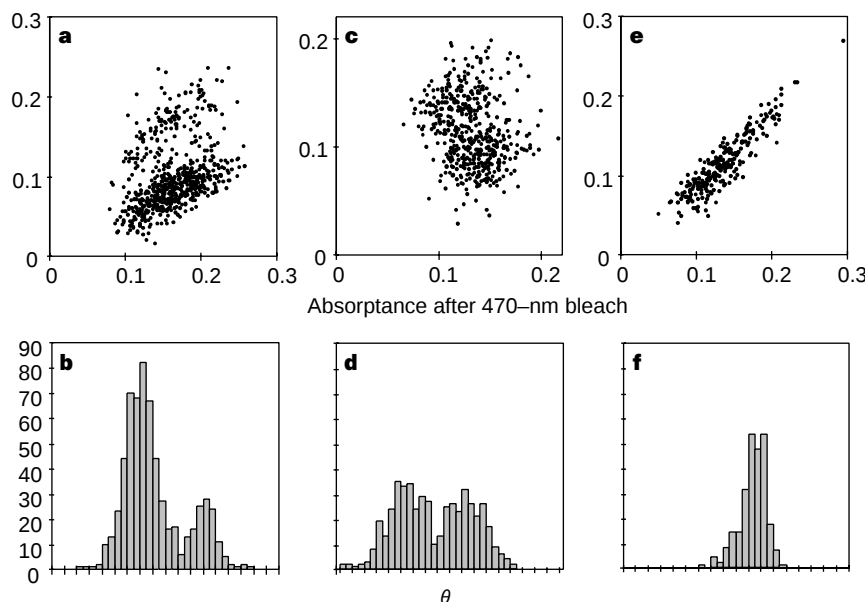


Figure 2 Scatter plots and histograms of individual cone absorbances. Scatter plots show the absorbance of each cone in a contiguous patch after the 470-nm and 650-nm selective bleaches. Cone absorbance was taken as the average value computed within a 0.4-arcmin square region centred on the cone. Histograms show the distribution of cones as a function of angle (θ) in the scatter plots. **a, b**, Results derived from the absorbance images of Fig. 1c, d, with S cones removed. **c, d**, Results obtained from a second trichromat, AN. For these trichromats, we fitted the sum of two gaussian curves to the histograms. The

angle corresponding to the intersection of the two gaussian curves was used to categorize L and M cones. The fractional area of the overlap compared with the total area under the two gaussians provides an estimate of the fraction of cones that were misidentified; these estimates are 2.1% for JW and 5.6% for AN. **e, f**, Results obtained from a protanope whose colour deficiency was verified psychophysically and by genetic screening. As with the trichromats, bleaching levels were chosen to optimize the chance of distinguishing two pigments.

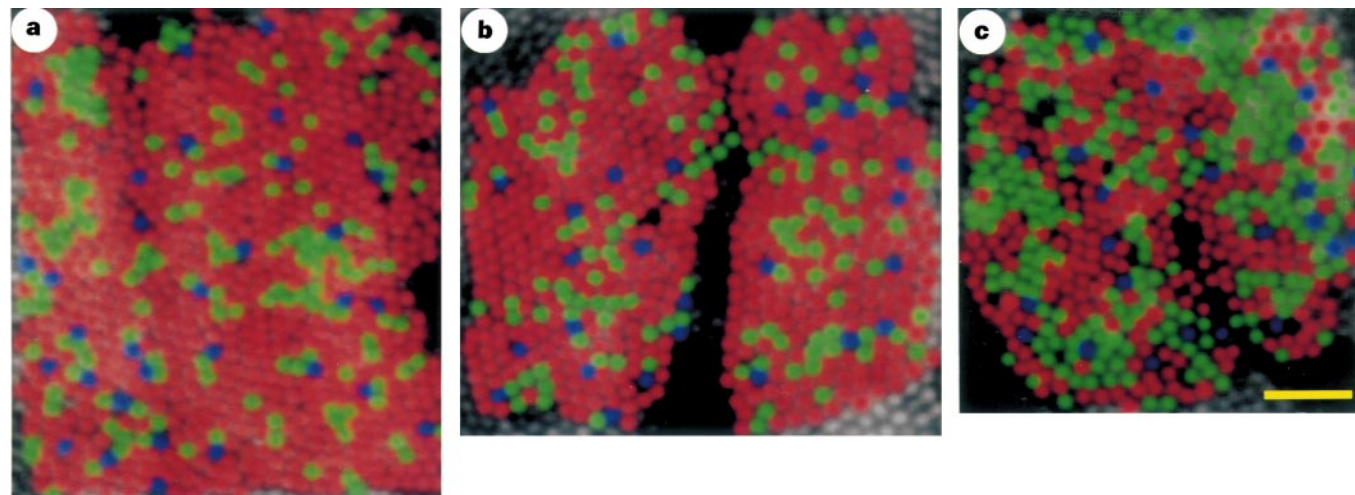


Figure 3 Pseudocolour image of the trichromatic cone mosaic. Blue, green and red colours represent the S, M and L cones, respectively. **a**, **b**, Subject JW's temporal and nasal retina, respectively, at one degree of eccentricity. **c**, Subject AN's nasal retina, at one degree of eccentricity. We performed a statistical test for randomness according to Diggle²⁷. We compared the distribution of all intercone distances of the measured M-cone array with 100 simulations derived from the

same mosaic in which the same number of M cones were randomly assigned. JW's array was no different from random at either location. AN's array showed significant clumping of the data ($P < 0.01$) but, because of optical blur, the possibility of a random assignment of M cones cannot be ruled out. Scale bar represents 5 arcmin of visual angle.

colour appearance of tiny, monochromatic light flashes^{19,20} because the relative excitation of different cone classes depends on the location of the flash. A related illusion is Brewster's colours, the perception of irregular patches of pastel colour while viewing periodic black and white patterns of high spatial frequency^{1,21}. Similarly, red-green isoluminant gratings with spatial frequencies above the resolution limit look like chromatic and luminant spatial noise²². All of these perceptual errors are examples of the aliasing produced when the three cone submosaics sample the retinal image inadequately. They are akin to the errors that occur in images taken with digital cameras that have interleaved pixels of different spectral sensitivity¹. The clumping that results from either the random or the aggregated assignment of M and L cones exacerbates these errors.

Both of our subjects have retinal patches of 5 arcmin or more across that contain only one of the two longer-wavelength-sensitive cone classes. Although the existence of these patches indicates that the trichromat may sometimes misjudge the colour appearance of tiny objects, the patches will be beneficial in recovering high-frequency luminance patterns, because cortical neurons tuned to high spatial frequency are more likely to be fed by contiguous cones of the same class. This may explain the observation that, in some normal trichromats, there is little or no difference in resolution for gratings seen with only M cones or only L cones, or when both cone classes operate together²³. Only when one cone class is greatly under-represented, as occurs in some heterozygous carriers for congenital X-linked protanopia, is resolution clearly mediated by the more dense submosaic²⁴.

The large individual differences in numbers and arrangement of cone classes that we have observed indicate that evolution has not converged on an optimum proportion of M and L cones for the human eye. Is this because red-green colour vision is a relatively new feature of vision in old-world primates^{25,26}, or do the statistics of natural scenes, optical blurring and clever post-receptor processing make M- and L-cone topography unimportant for visual performance? The imaging method described here allows us to address these questions, because it is now possible to assess visual performance and the circuitry of the retina in eyes for which the trichromatic mosaic is known. □

Received 16 October; accepted 26 November 1998.

1. Williams, D. R., Sekiguchi, N., Haake, W., Brainard, D. H. & Packer, O. S. in *From Pigments to Perception*. (eds Valberg, A. & Lee, B. B.) 11–22 (Plenum, New York, 1991).

2. Williams, D. R., MacLeod, D. I. A. & Hayhoe, M. Punctate sensitivity of the blue sensitive mechanism. *Vision Res.* **21**, 1357–1375 (1981).
3. Curcio, C. A. *et al.* Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *J. Comp. Neurol.* **312**, 610–624 (1991).
4. Babcock, H. W. The possibility of compensating astronomical seeing. *Publ. Astron. Soc. Pacif.* **65**, 229–236 (1953).
5. Liang, J., Williams, D. R. & Miller, D. T. Supernormal vision and high-resolution retinal imaging through adaptive optics. *J. Opt. Soc. Am. A* **14**, 2884–2892 (1997).
6. Campbell, F. W. & Rushton, W. A. H. Measurement of the scotopic pigment in the living human eye. *J. Physiol. (Lond.)* **130**, 131–147 (1955).
7. Rushton, W. A. H. & Baker, H. D. Red/green sensitivity in normal vision. *Vision Res.* **4**, 75–85 (1964).
8. Pokorny, J., Smith, V. C. & Wesner, M. in *From Pigments to Perception* (eds Valberg, A. & Lee, B. B.) 23–34 (Plenum, New York, 1991).
9. Cicerone, C. M. & Nerger, J. L. The relative numbers of long-wavelength-sensitive to middle-wavelength-sensitive cones in the human fovea. *Vision Res.* **26**, 115–128 (1989).
10. Vimal, R. L. P., Pokorny, J., Smith, V. C. & Shevell, S. K. Foveal cone thresholds. *Vision Res.* **29**, 61–78 (1989).
11. Jacobs, G. H. & Neitz, J. in *Colour Vision Deficiencies* Vol. XI (ed. Drum, B.) 107–112 (Kluwer Academic, Netherlands, 1993).
12. Jacobs, G. H. & Deegan, J. F. Spectral sensitivity of macaque monkeys measured with ERG flicker photometry. *Vis. Neurosci.* **14**, 921–928 (1997).
13. Bowmaker, J. K. & Dartnall, H. J. A. Visual pigments of rods and cones in a human retina. *J. Physiol. (Lond.)* **298**, 501–511 (1980).
14. Dartnall, H. J. A., Bowmaker, J. K. & Mollon, J. D. Human visual pigments: microspectrophotometric results from the eyes of seven persons. *Proc. R. Soc. Lond. B* **220**, 115–130 (1983).
15. Yamaguchi, T., Motulsky, A. G. & Deeb, S. S. Visual pigment gene structure and expression in human retinae. *Hum. Mol. Genet.* **6**, 981–990 (1998).
16. Hagstrom, S. A., Neitz, J. & Neitz, M. Variation in cone populations for red-green color vision examined by analysis of mRNA. *NeuroReport* **9**, 1963–1967 (1998).
17. Mollon, J. D. & Bowmaker, J. K. The spatial arrangement of cones in the primate fovea. *Nature* **360**, 677–679 (1992).
18. Packer, O. S., Williams, D. R. & Bensinger, D. G. Photoreceptor transmittance imaging of the primate photoreceptor mosaic. *J. Neurosci.* **16**, 2251–2260 (1996).
19. Holmgren, E. Über den Farbensinn. *Comp. Rend. Congr. Period. Intern. Sci. Med. Copenhagen* **1**, 80–98 (1884).
20. Krauskopf, J. Color appearance of small stimuli and the spatial distribution of color receptors. *J. Opt. Soc. Am.* **54**, 1171 (1964).
21. Brewster, D. On the undulations excited in the retina by the action of luminous points and lines. *Lond. Edinb. Philos. Mag. J. Sci.* **1**, 169–174 (1832).
22. Sekiguchi, N., Williams, D. R. & Brainard, D. H. Efficiency in detection of isoluminant and isochromatic interference fringes. *J. Opt. Soc. Am. A* **10**, 2118–2133 (1993).
23. Williams, D. R. in *Advances in Photoreception: Proc. Symp. Frontiers Visual Sci.* 135–148 (National Academy, Washington DC, 1990).
24. Miyahara, E., Pokorny, J., Smith, V. C., Baron, R. & Baron, E. Color vision in two observers with highly biased LWS/MWS cone ratios. *Vision Res.* **38**, 601–612 (1998).
25. Nathans, J., Thomas, D. & Hogeness, D. S. Molecular genetics of human colour vision: the genes encoding blue, green and red pigments. *Science* **232**, 193–202 (1986).
26. Mollon, J. D. "Tho' she kneeled in that place where they grew...": The uses and origins of primate colour vision. *J. Exp. Biol.* **146**, 21–38 (1989).
27. Diggle, P. J. *Statistical Analysis of Spatial Point Patterns* (Academic, London, 1983).

Acknowledgements. We thank D. Brainard, D. Dacey, J. Jacobs, J. Liang, D. Miller and O. Packer for their assistance. We acknowledge financial support from the Fight for Sight research division of Prevent Blindness America (to A.R.) and the National Eye Institute and Research to Prevent Blindness (to D.R.W.).

Correspondence and requests for materials should be addressed to A.R. (e-mail: aroorda@popmail.opt.uh.edu).

Asymmetric Notch activation specifies photoreceptors R3 and R4 and planar polarity in the *Drosophila* eye

Manolis Fanto & Marek Mlodzik

Developmental Biology Programme, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Planar polarity is seen in epidermally derived structures throughout the animal kingdom^{1,2}. In the *Drosophila* eye, planar polarity is reflected in the mirror-symmetric arrangement of ommatidia (eye units) across the dorsoventral midline or equator; ommatidia on the dorsal and ventral sides of the equator exhibit opposite chirality^{3–5}. Photoreceptors R3 and R4 are essential in the establishment of the polarity of ommatidia^{6–11}. The R3 cell is thought to receive the polarizing signal, through the receptor Frizzled (Fz), before or at higher levels than the R4 cell, generating a difference between neighbouring R3 and R4 cells^{6,7,9,10}. Both loss-of-function and overexpression of Fz in the R3/R4 pair result in polarity defects and loss of mirror-image symmetry^{6,7,9,10,12}. Here we identify *Notch* and *Delta* (*Dl*) as dominant enhancers of the phenotypes produced by overexpression of *fz* and *dishevelled* (*dsh*), which encodes a signalling component downstream of Fz, and we show that *Dl*-mediated activation of Notch is required for establishment of ommatidial polarity. Whereas *fz* signalling is required to specify R3, *Notch* signalling induces the R4 fate. Our data indicate that *Dl* is a transcriptional target of Fz/Dsh signalling in R3, and activates Notch in the neighbouring R4 precursor. This two-tiered mechanism explains how small differences in the level and/or timing of Fz activation reliably generate a binary cell-fate decision, leading to specification of R3 and R4 and ommatidial chirality.

Epithelial planar polarity (EPP) in the *Drosophila* eye is generated when immature ommatidial preclusters rotate by 90° towards the equator and adopt opposite chiral forms on either side of the equator^{3,5}. In mature ommatidia, the R3 and R4 photoreceptors are asymmetrically positioned at the tip of the ommatidial trapezoid, giving the clusters a chirality (Fig. 1). In the eye disc, initially R3 is equatorial and R4 is polar. In *fz* and *dsh* mutants, both chirality and rotation become random^{6,12,13}. In addition, in these mutants some ommatidia remain symmetrical, containing non-chiral R3/R3 or R4/R4 pairs instead of the chiral R3/R4 pair (Fig. 1b), indicating that the assignment of the R3 and R4 fates may be critical for the generation of chirality. Fz, the founding member of a family of seven-transmembrane-domain receptor proteins for the Wnt¹⁴ growth factors, and Dsh, a cytoplasmic signalling molecule that acts downstream of Fz, are part of the EPP signalling cascade^{7,9,13,15,16} that is thought to interpret a polarizing signal emanating from the equator. *fz* is required in R3 for EPP generation⁶. The polarizing signal is thought to be received first, or at higher levels, by the equatorial R3 cell and this generates a difference between the two precursor cells and leads to chirality. Accordingly indiscriminate activation of Fz signalling by overexpression of either Fz or Dsh in both R3 and R4 (using the *sev* enhancer, for example, *sev-Fz* and *sev-Dsh*) overrides the polarity-signal-induced difference, polarity again becomes random, and many clusters remain symmetrical (Fig. 1c; see also refs 7, 9, 10). As the R3/R4 precursors lie close together, an interesting question is how a small difference in timing or level of Fz/Dsh signalling could reliably induce the binary R3/R4-fate decision.

Using a genetic approach to identify other components of EPP

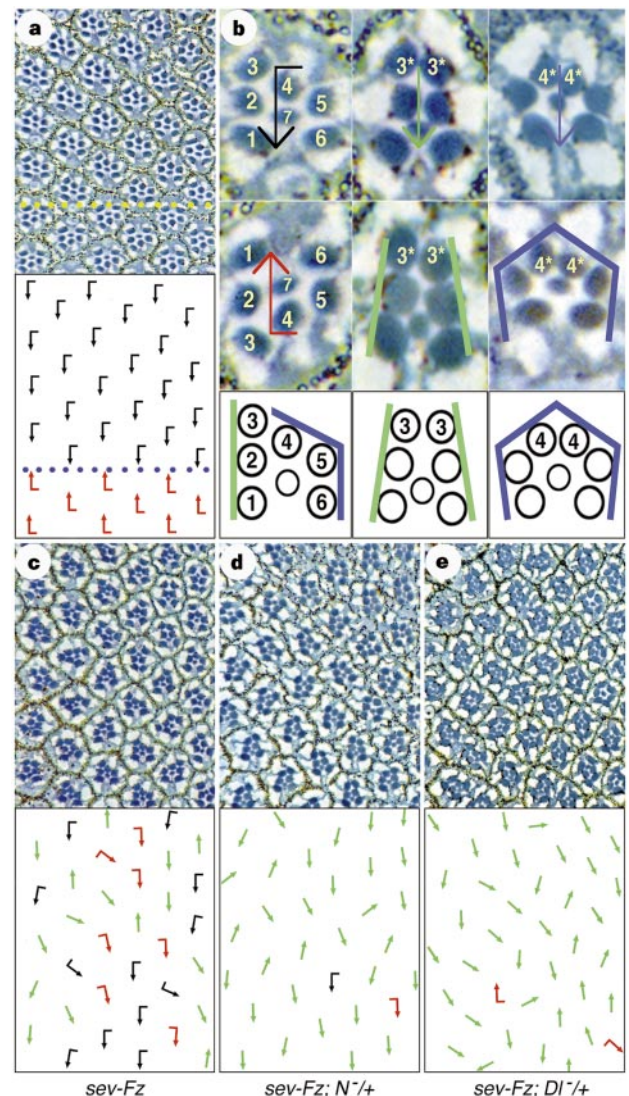


Figure 1 *Dl* and *Notch* dominantly enhance an eye-polarity-specific *fz* gain-of-function phenotype. All panels show tangential sections of dorsal regions (except **a** which shows an equatorial region of adult eyes). Anterior is left and the equator down in all figures. Each panel shows an eye section (top) and a corresponding diagram with arrows reflecting ommatidial polarity. How arrows relate to actual cell arrangement and polarity is shown in **b**. **a**, Wild-type equatorial region (equator is indicated by yellow and blue dotted lines in top and bottom panels, respectively), showing ommatidia of dorsal and ventral chirality (black and red arrows, respectively, in bottom panel). **b**, Ommatidial polarity described by arrows. Left, a wild-type dorsal ommatidium (top, black arrow) and ventral ommatidium (middle, red arrow) (photoreceptors R1–R7 are numbered), and an asymmetric representation of the wild-type ommatidium (bottom). Left and centre, symmetrical ommatidia of the R3/R3 (green arrow) and the R4/R4 (blue arrow) types. The R3/R3 type is identified by the symmetrical arrangement of green lines, representing the anterior R3 axis within a wild-type ommatidium (compare with **b**). The R4/R4 type is identified by the symmetry of the wild-type posterior R4 axis (blue), and can thus easily be distinguished from the R3/R3 type. **c**, *sev-Fz*/+, **d**, *sev-Fz*/+; *N*²⁶⁴⁻³⁹/+, and **e**, *sev-Fz*/+; *Dl*^{revF10}/+ ommatidia. The disorganized appearance caused by the autonomous overexpression of Fz signalling in the *sev-Fz* background is dominantly enhanced by mutations in *Dl* and *Notch*. Note that in *Dl* and *Notch* heterozygous backgrounds, most ommatidia show the R3/R3-type symmetry (green arrows). Whereas *sev-Fz*; +/+ mutants have 64.3 ± 7.4% (*n* = 815) R3/R3-type ommatidia, *sev-Fz*; *N*²⁶⁴⁻³⁹/+ and *sev-Fz*; *Dl*^{revF10}/+ flies have 89.8 ± 4.3% and 79.4 ± 5.4% R3/R3 symmetrical ommatidia respectively. Several mutant alleles of both *Notch* and *Dl* were tested and quantified, and show comparable interactions. *Serrate* does not show an interaction with *Sev-Fz* in this assay (*sev-Fz*; *Ser*^{FLX}/+ had 56.2 ± 9.2% symmetrical clusters). Similar interactions were observed between *sev-Fz* and the *Notch* and *Dl* alleles (not shown).

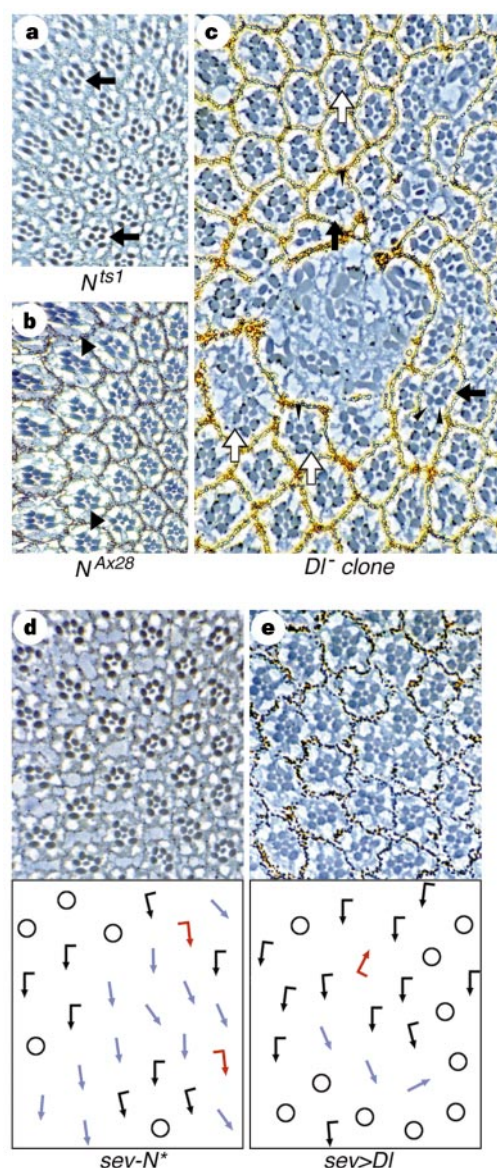


Figure 2 Notch signalling is required for the establishment of ommatidial chirality. Tangential sections of dorsal regions of adult eyes (**a–e**) and the corresponding representation (**d, e**) are shown. **a**, *N^{ts1}* mutant raised at 25 °C. **b**, *N^{Ax28}* mutant. Black arrows indicate R3/R3-type ommatidia and arrowheads the R4/R4-type. **c**, *Df^{evF10}* clone (marked by absence of pigment as detected in pigment cells, yellow granules, and black dots next to the rhabdomeres in photoreceptor cells). The margin of the clone that contains mosaic ommatidia is informative in the context analysed (the typical *DI*-mutant phenotype, with an excess of photoreceptors, is seen in the centre of the clone). The wild-type dorsal arrangement is evident in the pigmented area (examples are highlighted by white arrows). In *DI* mosaic R3/R4 pairs, the *DI*-positive cell always adopted the R3 fate. This can lead to polarity reversals and inverted chirality (black vertical arrow). The unpigmented cells within mosaic R3/R4 pairs are marked by black arrowheads. White arrows indicate the wild-type arrangement in completely wild-type or in mosaic R3/R4 pairs. The black horizontal arrow indicates an R3/R3-type ommatidium in which both cells of the pair are *DI* mutant (black arrowheads). **d**, *sev-Notch^{Δecd}* and **e**, *sev>DI* ommatidia. Note that a reduction in Notch activity (shown in **a**) leads to the development of symmetrical R3/R3-type ommatidia, whereas increased or indiscriminate Notch activation within R3/R4 leads to R4/R4-type symmetrical ommatidia (**b–e**; shown by blue arrows in **d, e**). Activation of Notch can also cause loss and transformation of outer photoreceptors R1/R6. Some of these ommatidia cannot reliably be scored for polarity (represented by circles in **d, e**).

signalling pathways, we screened for second-site modifiers of *sev-Fz* and *sev-Dsh* (refs 7, 9). In parallel, we tested components of known signalling pathways for genetic interactions with these genotypes (refs 7, 9 and M.F. and M.M., unpublished observations). This approach identified *Delta* (*DI*) and *Notch* (*N*) as dominant enhancers of *sev-Fz* and *sev-Dsh* (Fig. 1d, e; in contrast to positively acting components of the Fz/Dsh cascade, which act as suppressors^{7,9}). Reducing either *DI* or *Notch* function in these backgrounds increased the percentage of symmetrical ommatidia containing an R3/R3 photoreceptor pair in the normal R3/R4 positions (Fig. 1c–e). In contrast, reduction of the dosage of *Serrate*, which encodes another Notch ligand¹⁷, had no effect (Fig. 1 legend). These observations indicate that *DI* and Notch signalling may be involved in the generation of R3/R4 asymmetry and chirality.

To confirm that *Notch* is involved in the establishment of R3/R4 polarity, we tested whether loss of *Notch* function alone can produce

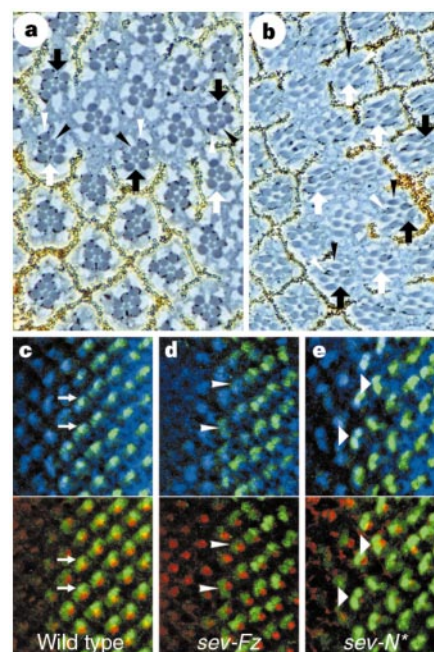


Figure 3 Fz and Notch signalling are sufficient to autonomously induce the R3 and R4 fates, respectively. **a, b**, Tangential sections of **a**, a *sev-Fz* and **b**, a *sev-N⁺* clone, respectively, in an otherwise wild-type background. The *sev-Fz* or *sev-N⁺*-containing cells are marked by the presence of pigment granules (black dots next to the rhabdomeres in photoreceptor cells). The wild-type dorsal arrangement is seen in the unpigmented clones (examples are highlighted by white arrows). **a**, In *sev-Fz*-mosaic R3/R4 pairs, the *sev-Fz*-positive cell always adopted the R3 fate. This can lead to polarity reversals and inverted chirality (black arrows). **b**, In *sev-N⁺*-mosaic R3/R4 pairs, the *sev-N⁺*-positive cell always adopted the R4 fate; again this can lead to polarity reversals (black arrows). White and black arrowheads indicated pigmented and unpigmented cells in the R3/R4 pair. **c–e**, Confocal images of triple-labelled H123-expressing third-instar eye disc in the following backgrounds: **c**, wild-type, **d**, *sev-Fz*, and **e**, *sev-N⁺*. Green, H123 staining (anti-β-galactosidase); blue: anti-Spalt, marking R3/R4 precursors transiently; red: rhodamine-phalloidin, highlighting membrane-associated actin and central membrane tufts of ommatidial preclusters; cells staining both green and blue appear turquoise. Upper panels, blue and green channels; lower panels, red and green channels. The morphogenetic furrow is at the left margin of each panel. In wild-type ommatidia, H123 is expressed at high levels in R4 (examples marked with arrows) and at low levels in R3. An overlap with Spalt expression of about one to two columns. More posteriorly Spalt is expressed in many cells but not in R3/R4. In *sev-Fz* discs, H123 expression is often low in both cells of the pair, reflecting an R3/R3 ommatidium (examples marked by thin arrowheads), whereas in *sev-N⁺* discs most clusters show high expression in both cells (examples marked by broad arrowheads), reflecting the R4/R4 type.

an eye phenotype similar to that generated by *sev-Fz*. The Notch pathway has been implicated in many cell–cell communication events. *Notch* loss-of-function mutations affect aspects of eye development ranging from proliferation and early patterning to photoreceptor differentiation^{18–24}. Thus, *Notch* mutants exhibit very pleiotropic phenotypes (data not shown). However, the temperature-sensitive *Notch* allele (*N^{ts1}*) is functional at low temperatures (18–22 °C) and lethal at 29 °C. We raised *N^{ts1}* larvae at roughly the threshold temperature (25 °C), thus barely reducing *Notch* function and allowing most *Notch*-mediated events to take place normally. We analysed the eyes of these flies for potential mild polarity defects. Although ommatidia in such eyes always contained the normal complement of photoreceptors, they sometimes remained symmetrical, exhibiting an R3/R3-like phenotype (Fig. 2a; compare with Fig. 1b). In contrast, eyes of the gain-of-function *Notch* allele *N^{Ac28}* showed mild polarity defects in which ommatidia had an R4/R4 phenotype (Fig. 2b), opposite to that produced by loss of *Notch* function.

To determine the role of *Dl* in R3/R4 polarity generation, we analysed *Dl*-mutant clones (Fig. 2c). To avoid effects on equator establishment^{22–24} we analysed cones that were located away from the dorsoventral midline. Although most of the mutant area showed the typical *Dl*-mutant phenotype of an excess of photoreceptors, the nature of the margins of the clone that contains mosaic clusters was informative. In ommatidia mosaic for *Dl* in R3/R4 photoreceptors, the *Dl⁺* cell always adopted the R3 fate. This could lead to polarity reversals and inverted chirality (black vertical arrow in Fig. 2c). Ommatidia with a normal number of photoreceptors in which both cells of the R3/R4 pair were *Dl⁺* showed an R3/R3 phenotype (black horizontal arrow in Fig. 2c), similar to those seen in *Notch* loss-of-function mutants (Fig. 2a).

To determine whether Notch functions cell-autonomously in the R3/R4 pair in this context, we expressed the Notch ligand *Dl* and an activated Notch in the R3 and R4 precursors (*sev > Dl* and *sev-N^{Δecd}*, referred to here as *sev-N^{*}*; ref. 19; see Methods). These genotypes result in many symmetrical ommatidial clusters exhibiting the R4/R4 phenotype (Fig. 2d, e; in addition to polarity, *sev-N^{*}* and *sev > Dl* affect outer photoreceptor R1/R6 development, ref. 19 and data not shown). The *sev-N^{*}* phenotype was not affected by reduction of dosage of *fz* and *dsh* (data not shown). Moreover, expression of *sev-N^{*}* in the *dsh¹* mutant background resulted in a phenotype that was identical to that produced by *sev-N^{*}* in a wild-type background (Fig. 2c and data not shown), indicating that *Notch* is epistatic to Fz/Dsh signalling in this context. Taken

together, these data indicate that, downstream of Fz, *Dl* is required in R3 cells and *Notch* signalling is required to induce the R4 fate.

Loss-of-function analysis has shown that *fz* is required in R3 to establish ommatidial polarity⁶ and our data indicate that *Notch* is required in R4 (Fig. 2). To test whether activation of Fz and Notch is sufficient to specify the R3 and R4 fates, respectively, we analysed the effects of the gain-of-function *sev-Fz* and *sev-N^{*}* mutants in mosaic ommatidia in an otherwise wild-type background. This approach revealed that, in the case of Fz, whenever there was mosaicism within the R3/R4 pair, the *sev-Fz*-positive cell adopted the R3 fate (and the other cell adopted the R4 fate; Fig. 3a). The same analysis with *sev-N^{*}* showed the opposite effect, that is, the *sev-N^{*}*-positive cell always adopted the R4 fate (Fig. 3b). Thus, high levels of Fz signalling induce the R3 fate, cell-autonomously, whereas full activation of Notch induces R4.

To confirm these observations at the earliest detectable stage, we used the enhancer-trap line H123, in which β-galactosidase is expressed at high levels in R4 and at lower levels in R3 (Fig. 3b). In polarity mutants such as *fz*, *dsh* and *strabismus*, the differential expression of β-galactosidase within the R3/R4 pair is randomized in accordance with the adult phenotype of random chirality¹¹. We tested whether the R3/R4 fate is affected by *sev-Fz* and *sev-N^{*}* signalling by analysing H123 expression in imaginal discs (Fig. 3c, d). In *sev-Fz* discs, H123 staining within the R3/R4 pair was either random or low in both cells of the pair, indicating that both cells were R3 cells (Fig. 3d). In contrast, activated Notch (*sev-N^{*}*) led mainly to high levels of H123 expression in both cells of most R3/R4 pairs, indicating that both cells were R4 cells (Fig. 3e). These data show that Fz specifies the R3 cell fate and Notch signalling induces the R4 fate.

The *Dl* protein is expressed in the R3 precursor at high levels during the period of EPP generation²⁵. Together with our analysis of *Dl*-mutant clones (Fig. 2c), this result indicates that activation of Notch signalling in R4 may be achieved by upregulation of *Dl* in the R3 precursor. *Dl* might be transcriptional target of Fz/Dsh activation in R3 cells. We therefore monitored the expression pattern of *Dl* in R3/R4 precursors of wild-type and *sev-Fz* discs using a *Dl* enhancer-trap line as a marker (Fig. 4). In wild-type discs *Dl* was expressed in a dynamic pattern in many cells; this pattern included transient expression of *Dl* in the R3 precursor (Fig. 4a). Later, expression was downregulated in this R3 precursor as the cluster matured (Fig. 4a). In a *sev-Fz* background, expression of the *Dl* enhancer-trap line was upregulated in both cells of the R3/R4-precursor pair and this expression was maintained in several

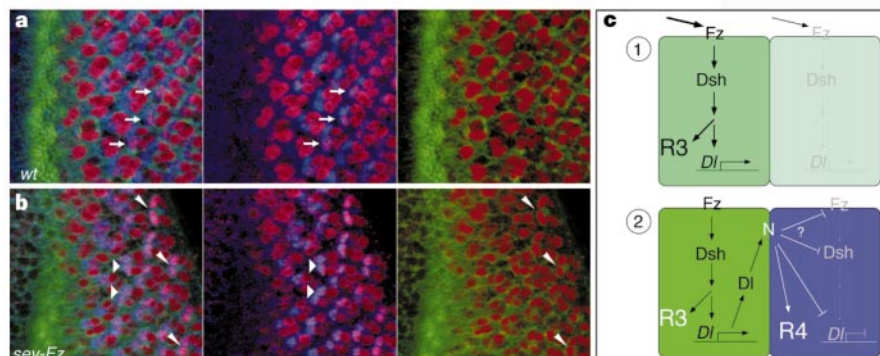


Figure 4 *Dl* expression in the R3/R4 pair is regulated by Fz signalling. **a**, **b**, *Dl*-enhancer-trap line expression **a**, a wild-type and **b**, a *sev-Fz* disc. Regions shown are close to the morphogenetic furrow. Red, β-galactosidase expression in the *Dl*-enhancer-trap line; blue, anti-Spalt staining, highlighting R3/R4 cells; green, phalloidin staining to highlight actin, morphogenetic furrow and cluster geometry. Cells co-expressing Spalt and *Dl* appear purple in the overlay. Left, triple staining, all channels; middle, red and blue channels; right, green and red channels. **a**, Note

that *Dl* expression is restricted to R3 cells and is transient (some examples are indicated with arrows). **b**, Both cells of the R3/R4 pair express *Dl* and appear purple in the left and middle panels (some examples are indicated with arrowheads); they also maintain *Dl* expression for several hours, coinciding with *sev-Fz* expression. **c**, A two-step model for R3/R4 specification and chirality generation.

columns in more mature clusters (Fig. 4b; compare the relative intensity of staining in R3/R4 and R1/R6 cells in older clusters, some of which are highlighted by arrowheads). Thus, *Dl* is a transcriptional target of Fz signalling in the eye disc.

Our data indicate that the interpretation of a planar polarity signal and the generation of R3/R4 asymmetry during eye development uses a two-tiered mechanism (Fig. 4c). First, activation of Fz in the equatorial cell of the R3/R4-precursor pair leads to a bias towards R3 in one of the cells, and *Dl* expression is upregulated in this cell. In the second step, induced *Dl* expression in R3 in turn activates Notch in the neighbouring R4 precursor, inducing the R4 fate and/or inhibiting Fz-mediated R3-fate induction in this cell. This locks the binary R3/R4-fate decision into place. This mechanism explains how a small difference in the timing or level of Fz activation in the R3/R4 pair generates a binary fate decision between these two cells. These data also show why even complete loss-of-function of either *fz* or *dsh* produces ommatidia that acquire some chirality, albeit at random^{6,9,10,13}. In these mutants it could be the *Dl*-Notch interaction that is responsible for chirality generation (by inducing R4). Without Fz signalling, *Dl* expression is not upregulated and the decision to follow an R3 or R4 fate is made in a stochastic manner. How Notch signalling induces the R4 fate remains unclear, as it usually represses photoreceptor development at this stage¹⁹. However, the precursor cells are already committed to form the R3/R4 pair by transcription factors such as Seven-up that are required for both R3/R4-cell fate²⁶ and polarity generation⁸.

Polarized Notch activation in response to signalling pathways such as that described here is likely to be of general significance. Notch signalling is also involved in the precise positioning of neural cells within proneural clusters, which is not random²⁷ (although all cells are competent to become neurons, it is not a stochastic decision), and polarized expression of Notch-pathway components has also been described in multicellular feather primordia in vertebrates²⁸. □

Methods

Fly strains and histology. The *sev-Fz*, *sev-Dsh* and *sev-N^{Δecd}* strains were as described^{7,9,19}. Flies were reared at 25 °C and, for study of genetic interactions, were heterozygous for both *sev-Fz* and the mutation of interest. The *sev > Dl* strain was generated with the Gal4 upstream activation sequence (UAS) system²⁹ from a *sev-GAL4* and a *UAS-Dl* line. Mutant alleles used were *dsh¹*, *fz¹*, *Dl¹*, *Dl^{rev10}*, *Dl^K*, *Dl^{PP}*, *N^{5e11}*, *N²⁶⁴⁻³⁹*, *N²⁶⁴⁻⁴⁰* and *Ser^{RXS}*. The *Dl* enhancer-trap line (*Dl-lac 1282*) was a gift from M. Haenlin. Adult eye sections and treatment of eye imaginal discs for antibody stainings was done as described³⁰.

Immunofluorescence. The following primary antibodies were used: rat-anti-Spalt (gift from R. Barrio), rabbit-anti-β-galactosidase (Cappel) and rhodamine-phalloidin (gift from A. Ephrussi). Secondary antibodies were conjugated with fluorescein isothiocyanate, rhodamine isothiocyanate or Cy5 (Jackson Labs). Following antibody incubations and washes, the eye imaginal discs were mounted in Mowiol and viewed with a Leica confocal microscope; images were assembled in Adobe Photoshop.

Received 20 October 1998; accepted 7 January 1999.

- Adler, P. N. The genetic control of tissue polarity in *Drosophila*. *BioEssays* **14**, 735–741 (1992).
- Eaton, S. Planar polarization in *Drosophila* and vertebrate epithelia. *Curr. Opin. Cell Biol.* **9**, 860–866 (1997).
- Tomlinson, A. Cellular interactions in the developing *Drosophila* eye. *Development* **104**, 183–193 (1988).
- Ready, D. F. A multifaceted approach to neural development. *Trends Neurosci.* **12**, 102–110 (1989).
- Wolff, T. & Ready, D. F. In *The Development of Drosophila melanogaster* (eds Martinez-Arias, A. & Bate, M.) 1277–1326 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1993).
- Zheng, L., Zhang, J. & Carthew, R. W. *frizzled* regulates mirror-symmetric pattern formation in the *Drosophila* eye. *Development* **121**, 3045–3055 (1995).
- Boutros, M., Paricio, N., Strutt, D. I. & Mlodzik, M. Dishevelled activates JNK and discriminates between JNK pathways in planar polarity and wingless signaling. *Cell* **94**, 109–118 (1998).
- Fanto, M., Mayes, C. A. & Mlodzik, M. Linking cell-fate specification to planar polarity: determination of the R3/R4 photoreceptors is a prerequisite for the interpretation of the Frizzled mediated polarity signal. *Mech. Dev.* **74**, 51–58 (1998).
- Strutt, D. I., Weber, U. & Mlodzik, M. The role of RhoA in tissue polarity and Frizzled signalling. *Nature* **387**, 292–295 (1997).
- Tomlinson, A., Strapps, W. R. & Heemskerk, J. Linking Frizzled and Wnt signaling in *Drosophila* development. *Development* **124**, 4515–4521 (1997).

- Wolff, T. & Rubin, G. M. *strabismus*, a novel gene that regulates tissue polarity and cell fate decisions in *Drosophila*. *Development* **125**, 1149–1159 (1998).
- Gubb, D. Genes controlling cellular polarity in *Drosophila*. *Development Suppl.*, 269–277 (1993).
- Theisen, H. et al. *dishevelled* is required during wingless signalling to establish both cell polarity and cell identity. *Development* **120**, 347–360 (1994).
- Bhanot, P. et al. A new member of the frizzled family from *Drosophila* functions as a Wingless receptor. *Nature* **382**, 225–230 (1996).
- Krasnow, R. E., Wong, L. L. & Adler, P. N. *dishevelled* is a component of the frizzled signaling pathway in *Drosophila*. *Development* **121**, 4095–4102 (1995).
- Axelrod, J. D., Miller, J. R., Shulman, J. M., Moon, R. T. & Perrimon, N. Differential recruitment of Dishevelled provides signaling specificity in the planar cell polarity and Wingless signaling pathways. *Genes Dev.* **12**, 2610–2622 (1998).
- Artavanis-Tsakonas, S., Matsuno, K. & Fortini, M. Notch signaling. *Science* **268**, 225–232 (1995).
- Cagan, R. L. & Ready, D. F. *Notch* is required for successive cell decisions in the developing *Drosophila* retina. *Genes Dev.* **3**, 1099–1112 (1989).
- Fortini, M. E., Simon, M. A. & Rubin, G. M. Signalling by the *sevenless* protein tyrosine kinase is mimicked by Ras1 activation. *Nature* **355**, 559–561 (1992).
- Baker, N. E. & Zitron, A. E. *Drosophila* eye development: Notch and Delta amplify a neurogenic pattern conferred on the morphogenetic furrow by *scabrous*. *Mech. Dev.* **49**, 173–189 (1995).
- Baker, N. E., Sung, Y. & Han, D. Evolution of proneural *atonal* expression during distinct regulatory phases in the developing *Drosophila* eye. *Curr. Biol.* **6**, 1290–1301 (1996).
- Papayannopoulos, V., Tomlinson, A., Panin, V. M., Rauskolb, C. & Irvine, K. D. Dorsal-ventral signaling in the *Drosophila* eye. *Science* **281**, 2031–2034 (1998).
- Cho, K.-O.K. & Choi, K. W. Fringe is essential for mirror symmetry and morphogenesis in the *Drosophila* eye. *Nature* **396**, 272–276 (1998).
- Dominguez, M. & de Celis, J. F. A dorsal/ventral boundary established by Notch controls growth and polarity in the *Drosophila* eye. *Nature* **396**, 276–278 (1998).
- Parks, A. L., Turner, F. R. & Muskavitch, M. A. T. Relationships between complex Delta expression and the specification of retinal cell fates during *Drosophila* eye development. *Mech. Dev.* **50**, 201–216 (1995).
- Mlodzik, M., Hiromi, Y., Weber, U., Goodman, C. S. & Rubin, G. M. The *Drosophila* *seven-up* gene, a member of the steroid receptor gene superfamily, controls photoreceptor cell fates. *Cell* **60**, 211–224 (1990).
- Campuzano, S. & Modolell, J. Patterning of the *Drosophila* nervous system: the achaete-scute gene complex. *Trends Genet.* **8**, 202–208 (1992).
- Chen, C. W. J., Jung, H. S., Jiang, T. X. & Chuong, C. M. Asymmetric expression of Notch/Delta/Serrate is associated with the anterior-posterior axis of feather buds. *Dev. Biol.* **188**, 181–188 (1997).
- Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
- Tomlinson, A. & Ready, D. F. Neuronal differentiation in the *Drosophila* ommatidium. *Dev. Biol.* **120**, 366–376 (1987).

Acknowledgements. We thank R. Barrio, M. Fortini, M. Haenlin, J. Royet, M. Milan, M. Muskavitch, A. Parks, G. M. Rubin and the Bloomington stock center for fly strains and antibodies; M. Fortini and S. Bray for sharing unpublished results and discussion; A. Cyrklaff for chromosome *in situ* mapping; C. Blummueller and J. Curtiss for comments on the manuscript; and members of the Mlodzik laboratory for stimulating discussions.

Correspondence and requests for materials should be addressed to M.M. (e-mail: mmlodzik@emblheidelberg.de).

Frizzled regulation of Notch signalling polarizes cell fate in the *Drosophila* eye

Michael T. D. Cooper & Sarah J. Bray

Department of Anatomy, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK

The *Drosophila* eye, a paradigm for epithelial organization, is highly polarized with mirror-image symmetry about the equator. The R3 and R4 photoreceptors in each ommatidium are vital in this polarity; they adopt asymmetrical positions in adult ommatidia and are the site of action for several essential genes^{1–5}. Two such genes are *frizzled* (*fz*) and *dishevelled* (*dsh*), the products of which are components of a signalling pathway required in R3, and which are thought to be activated by a diffusible signal^{3,6–10}. Here we show that the transmembrane receptor Notch is required downstream of *dsh* in R3/R4 for them to adopt distinct fates. By using an enhancer for the Notch target gene *Enhancer of split md*, we show that Notch becomes activated specifically in R4. We propose that Fz/Dsh promotes activity of the Notch ligand Delta and inhibits Notch receptor activity in R3, creating a difference in Notch signalling capacity between R3 and R4. Subsequent feedback in the Notch pathway ensures that this difference becomes amplified. This interplay between Fz/Dsh and Notch indicates that polarity is established through local comparisons between

two cells and explains how a signal from one position (for example, the equator in the eye) could be interpreted by all ommatidia in the field.

Each ommatidium in *Drosophila* contains eight photoreceptors, which are recruited sequentially early in eye development. The clusters formed are bilaterally symmetrical with three pairs of outer photoreceptors (R2/R5, R3/R4, R1/R6), the R3/R4 pair being the most anterior^{1,2} (Fig. 1a, b). However, the adult ommatidia are asymmetrical; the rhabdomeres form a trapezoid with R3 at the apex and are orientated to point away from the equator. This

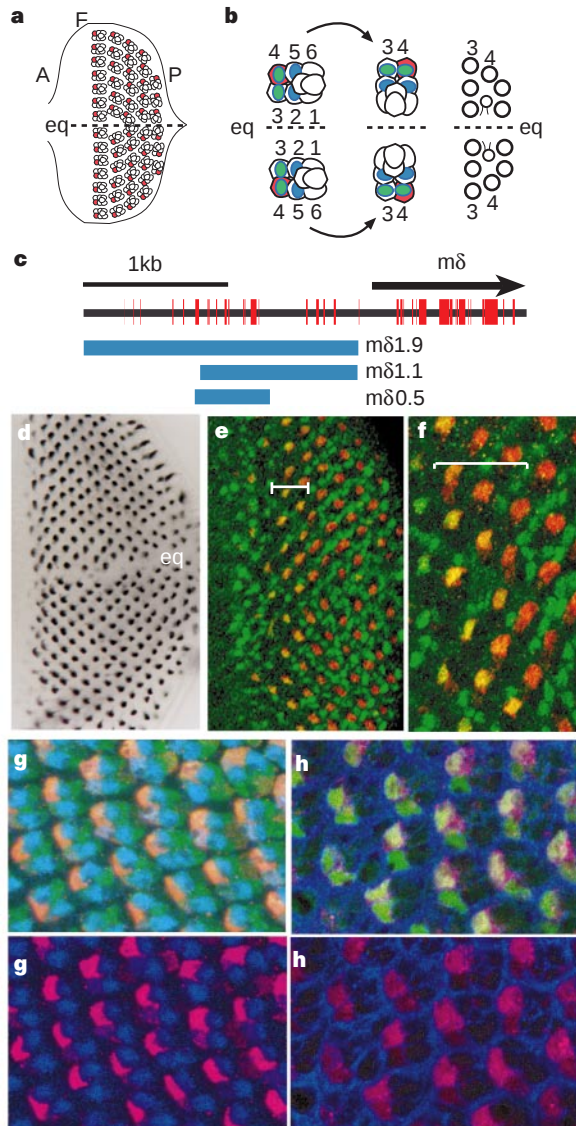


Figure 1 Specific expression of *E(spl)mδ* in R4. **a**, Diagram indicating ommatidial rotation in the eye imaginal disc. Equator (eq), furrow (F), anterior (A) and posterior (P); R4 is red. **b**, Orientation with respect to equator (eq) of unrotated eight-cell cluster, fully rotated eight-cell cluster and rhabdomeres in adult. Colours indicate Rough (blue) Spalt (green) and mδ0.5 (red). **c**, Enhancer fragments (blue) from *E(spl)mδ*. Red boxes are conserved in the *D. hydei* gene. Black arrow represents the transcript. **d-f**, Expression of mδ0.5 is detected in a single photoreceptor (**d**) and (**e**, **f**) overlaps with endogenous *E(spl)* (green) in rows 4-6 (red, mδ0.5). **g**, **g'**, **h**, **h'**, mδ0.5 is expressed in R4. Triple labelling to detect (**g**, **g'**) mδ0.5 (red), Rough (blue, strongest in R2 and R5, weaker in R3 and R4) and Elav (green, all photoreceptors) or (**h**, **h'**) mδ0.5 (red) Spalt (green, R3 and R4) and Coracle (blue, apical membranes) in rows ~5-9. **g'** and **h'**, red and blue channels only. Anterior is left, the equator below.

polarized organization requires differentiation between R3 and R4 and rotation of each cluster with respect to the equator (Fig. 1a, b). Defective Fz or Dsh activity produces ommatidia with random polarity^{3,6,7,9,10}. These proteins are required in R3, the cell closest to the equator in the R3/R4 pair, consistent with the hypothesis that Fz activity depends on a signal emanating from the equator^{3,10}. This indicates that the polarized behaviour of the ommatidia requires

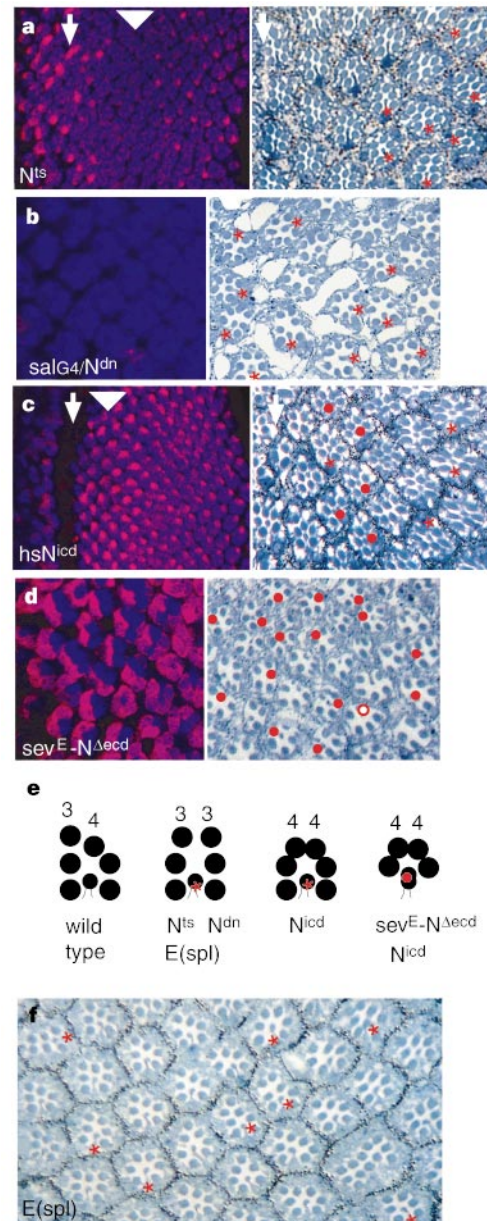


Figure 2 Notch activity is required to distinguish R3/R4 cell fates. **a-d**, Expression of mδ0.5 (red) and Elav (blue) and ommatidial phenotypes caused by reduced (**a**, **b**) or ectopic (**c**, **d**) Notch activity in the genotypes indicated. Arrowhead in **a**, **b** indicates region of modified mδ0.5 expression, adjacent (arrow) are regions where earlier stages of photoreceptor development have been affected (neural hypertrophy in *N^{ts}* (**a**); absence of neural development in *hsN^{icd}* (**b**); ref. 14). Symmetrical ommatidia are indicated by red asterisks or red solid circles, where the circles mark ommatidia with only four outer photoreceptors (open circle, asymmetrical ommatidium in **d**). As Spalt (R3/R4) but not Bar (R1/R6) are expressed in the *sev^{E-N^{Δecd}}*, we conclude the ommatidia are R4/R4/R2/R5 and not as previously reported¹⁷. **(e)** Summary of adult phenotypes. **(f)** Symmetrical ommatidia (asterisks) in eyes with reduced *E(spl)*; some ommatidia have extra photoreceptors¹⁸.

that a difference is established between the R3/R4 cell-pair before ommatidial rotation.

In analysing the regulation of *Enhancer of split mδ* (*E(spl)mδ*), we found specific expression in R4 at the time when neural antigens are first detected, which provides the earliest manifestation of a difference within the R3/R4 cell pair. Whereas larger *E(spl)mδ* fragments (mδ1.9, mδ1.1; Fig. 1c) confer a complex pattern that is reminiscent of the endogenous gene (data not shown), a ~500-base-pair (bp) fragment (mδ0.5; Fig. 1c) displays an R4-specific subset of the pattern (Fig. 1d–h). Although endogenous *E(spl)mδ* expression in this cell is transient, activity of mδ0.5 persists, indicating that it lacks an element needed for switching-off expression (Fig. 1e, f). Double-labelling of eye discs to detect proteins expressed in photoreceptor pairs confirmed that mδ0.5 is expressed specifically in the presumptive R4 (Fig. 1g, h).

E(spl) genes are targets of Notch signalling¹¹. The mδ0.5 fragment retains two binding sites for the nuclear Suppressor of Hairless

protein, a critical component of Notch signal transduction, which are responsive to Notch in tissue-cultured cells¹². To confirm that mδ0.5 expression depends on Notch activity *in vivo* we used two methods to modulate function in photoreceptor cells after recruitment, circumventing the early requirement for Notch in the same cells^{13,14}. First, using conditional alleles, we found that transiently reducing Notch activity for 6 h in late-third-instar larvae (using *N^{ts}*) leads to a loss of mδ0.5 expression from approximately four rows of ommatidia, whereas transient expression of a constitutively active Notch derivative (using *hs-N^{icd}*)¹⁵ has the converse effect (Fig. 2a, c). Second, expressing a dominant-negative form of Notch (*N^{dn}*)¹⁶ specifically in the photoreceptors eliminates mδ0.5 expression (Fig. 2), whereas expression of activated Notch (*N^{Δecd}*)¹⁷ results in ectopic mδ0.5 expression in R3 (and in additional photoreceptors in older ommatidia) (Fig. 2b, d). As expected, neither of these manipulations disturbs the equator itself (data not shown).

The R3 and R4 rhabdomeres are positioned asymmetrically in the

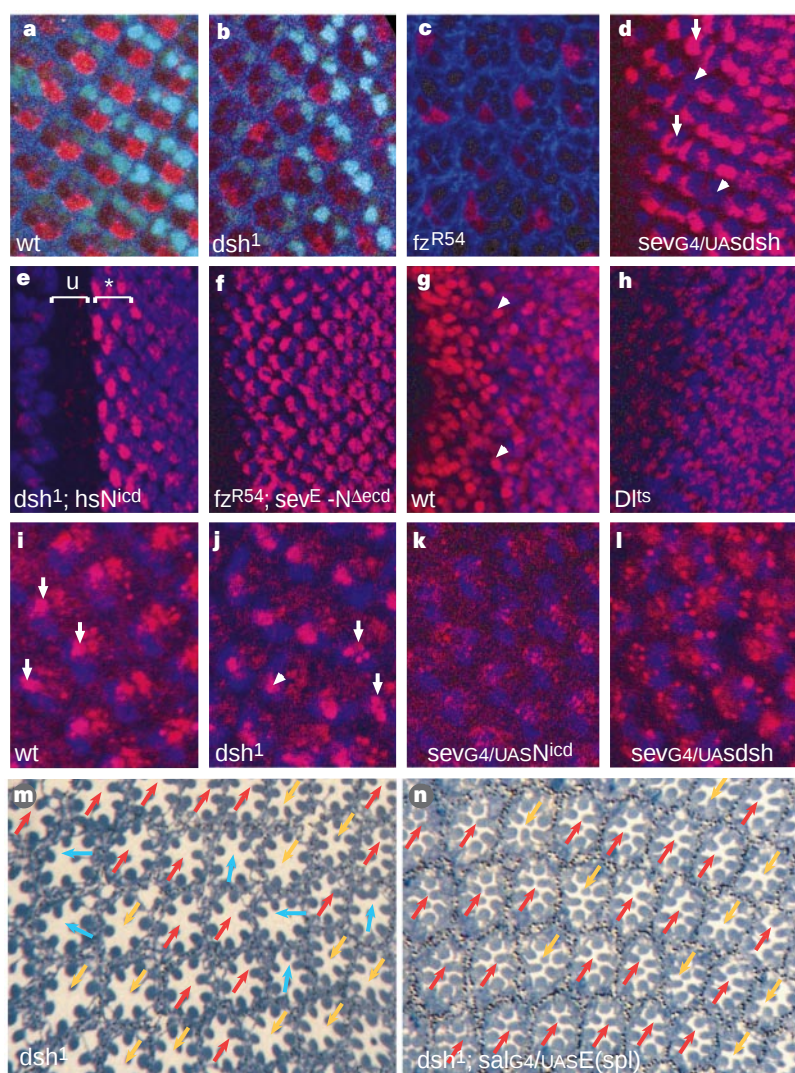


Figure 3 Notch activation is dependent on Fz/Dsh. **a–f**, Expression of mδ0.5 (red) is reduced in *dsh¹* (**b**) *fz^{R54}* (**c**) but elevated by Dsh overexpression (**d**) and can be rescued by expression of activated Notch (*hs-N^{icd}* or *sevE-N^{Δecd}*) in *dsh¹* (**e**) or *fz^{R54}* (**f**) discs. In **d**, examples where mδ0.5 is in both (arrow) or neither (arrowhead) R3/R4. In **e**, mδ0.5 (asterisk) is induced adjacent to a region where neural development is inhibited (u). Blue corresponds to anti-Coracle (**a–c**) and anti-Elav (**d–f**). Anti-Bar (turquoise in **a, b**) identifies R1/R6. Anterior is to the left and in **a–c** the equator is to the top. **g, h**, Transient reduction in *Df* (*Df^{ts}* exposed to

29°C for 3 h in **h**) leads to decreased Mδ (red) expression; arrowheads in **g** mark expression in R4, anti-Elav (blue). **i–l**, *Df* (red) expression in the genotypes indicated; anti-Spalt (blue) detects R3/R4 nuclei. In wild type (**i**), *Df* is highest in R3 (arrows), but in *dsh¹* (**j**), *Df* expression may be in both R3/R4 (arrows) or in the polar R4 position (arrowhead). *Df* levels (red) are reduced by *N^{icd}* (**k**) and increased by Dsh over-expression (**l**). **m, n**, Sections of adult retina from *dsh¹* (**m**) and from *dsh¹; sal-Gal4 UAS-E(spl)mδβ/+* (**n**); red arrows mark ommatidia with correct equatorial orientation, yellow misrotated by 180°, and turquoise other orientations.

adult ommatidia^{1,2}. If mδ0.5 expression reflects a requirement for Notch in specifying R4 fate we would expect loss or gain of Notch function to cause symmetrical ommatidia (that is, R3/R3 and R4/R4, respectively). Sections of adult eyes where Notch was manipulated as above confirm this prediction. In *N^{ts}* a patch of symmetrical ommatidia with the elongated shape expected for two R3 photoreceptors is generated (Fig. 2a), and induction of ectopic Notch activity (*hs-N^{icd}*) results in a patch of ommatidia with the converse R4/R4 organization (Fig. 2c). Misexpression of *N^{dm}* in R3/R4 results in >61% R3/R3 ommatidia ($\pm 11.3\%$ standard error; 10.5% asymmetrical, 28.5% unclassified; Fig. 2b) and in *sev^E-N^{Δecd}* > 69.5% of ommatidia are symmetrical ($\pm 8.8\%$; 12.2% asymmetrical, 18.3% unclassified; Fig. 2d), although ommatidia in *sev^E-N^{Δecd}* have only four outer photoreceptors (R4/R4/R2/R5 based on molecular markers). Symmetrical R3/R3 ommatidia (21%) are also present in flies with reduced *E(spl)* function (Fig. 2e and ref. 18). Thus both *Notch* and *E(spl)* are required to establish the R4 fate.

R3/R4 polarity within each ommatidium requires the Fz/Dsh pathway, which is mediated by the Dishevelled, Egl-10, Pleckstrin (DEP) domain in Dsh and does not involve the Wingless transduction cascade^{3,6-9}. If Fz activity regulates Notch signalling, mδ0.5 expression should be altered in *fz* and *dsh* mutants. As Dsh has been shown to antagonize Notch elsewhere¹⁹, we were surprised to find that mδ0.5 expression is markedly reduced in *dsh¹* (an allele that disrupts the DEP domain⁹) and *fz^{R54}* mutant larvae (Fig. 3a-c), which indicates that Fz/Dsh signalling promotes Notch activity in R3/R4. Expression of *N^{icd}* or *N^{Δecd}* in *fz* and *dsh* mutants restores mδ0.5 expression, confirming that Fz/Dsh are required upstream of Notch activation (Fig. 3e, f). The *dsh¹* protein has impaired signalling but the phenotype can be partially rescued by overexpressing downstream components⁷⁻⁹. If Notch activation is a key outcome of Dsh signalling, then overexpression of Notch targets might also rescue defects in *dsh¹* polarity. When *E(spl)* proteins are expressed in *dsh¹* eye discs (using *sevGal4*; Fig. 3m, n) the eyes are less roughened and 72.7% ($\pm 3.7\%$) of ommatidia have correct rotation and chirality compared with 55% ($\pm 4.9\%$) in *dsh¹*. *E(spl)* overexpression does not, however, product symmetrical ommatidia, which indicates that it is insufficient to confer R4 fate and is more likely to alter the threshold between the two fates by antagonizing R3 promoting genes. Additional targets of Notch needed to specify R4 identity may include *strabismus*⁵.

Our results show that Fz/Dsh signalling promotes Notch activation in R4 and, because Fz is required in R3, indicates that regulation could act through a Notch ligand such as Delta (Dl). Consistent with this, Dl levels are highest in R3 in wild-type ommatidia (Fig. 3i and ref. 20) and Mδ expression is greatly reduced in *Dl^{ts}* eye discs (Fig. 3g, h). In addition, overexpression of Dsh results in higher, more diffuse Dl expression (Fig. 3l). However, Dl is still present in *dsh¹* mutants, even though Notch signalling is compromised, but appears more equally distributed between R3/R4 (Fig. 3j). This indicates that the relative levels of Dl in R3 and R4 are important rather than the absolute levels and that there is also a mechanism for repressing *Dl* in the wild-type R4. This could involve Notch itself, downregulating ligand expression in the receiving cell as has been proposed (see ref. 21). The reduced Dl levels detected when *N^{icd}* is expressed supports this feedback (Fig. 3k). Thus we propose that signalling through Fz/Dsh results initially in higher Dl activity in the pre-R3 cell, which leads to a net activation of Notch in the adjacent photoreceptor. As a consequence, expression of *Dl* is inhibited in this adjacent cell, so further polarizing the difference in signalling capability between the two cells (Fig. 4d). The effects on Dl may be transcriptional as similar effects are seen using a *lacZ* gene inserted at the *Dl* locus (data not shown). If this model is correct, manipulations that decrease the difference in Fz/Dsh activity between R3 and R4 would remove the differential in Dl levels and should result in a more stochastic outcome. This occurs in eyes with reduced *fz/dsh*, where any residual mδ0.5 expression is randomized

between the R3/R4 pair (Fig. 3b, c), and in eyes with Dsh overexpressed, where mδ0.5 is detected in one, both or neither R3/R4 cell (Fig. 3d). In both cases the stochastic outcome correlates with the adult phenotypes^{3,6,8,9}.

Ommatidial polarity also requires *spiney-legs (sple)*, which seems to affect the early distinction between R3 and R4 (refs 22, 23). To test whether *sple* influences Notch signalling between R3/R4 we examined mδ0.5 expression in *sple¹*. The results indicate that *Sple* is normally inhibitory to Notch in R3 as we detect high levels of mδ0.5 in both cells of the R3/R4 pair for a prolonged period in *sple¹* mutant discs (Fig. 4a-c). Subsequently, mδ0.5 becomes restricted to one cell, but whereas normally it is always expressed in the polar R4 cell, in *sple¹* it refines to either cell without the normal bias (Fig. 4c). Thus, *Sple* affects the resolution of Notch signalling, rather than its initiation, apparently by making the presumptive R3 more resistant to Notch activation, a feature that is consistent with mosaic analysis indicating that *Sple* is required in R3 (ref. 23). If this antagonistic effect of *sple* is generally applicable then phenotypes caused by weak Notch mutations ought to be suppressed by mutations at the *sple*

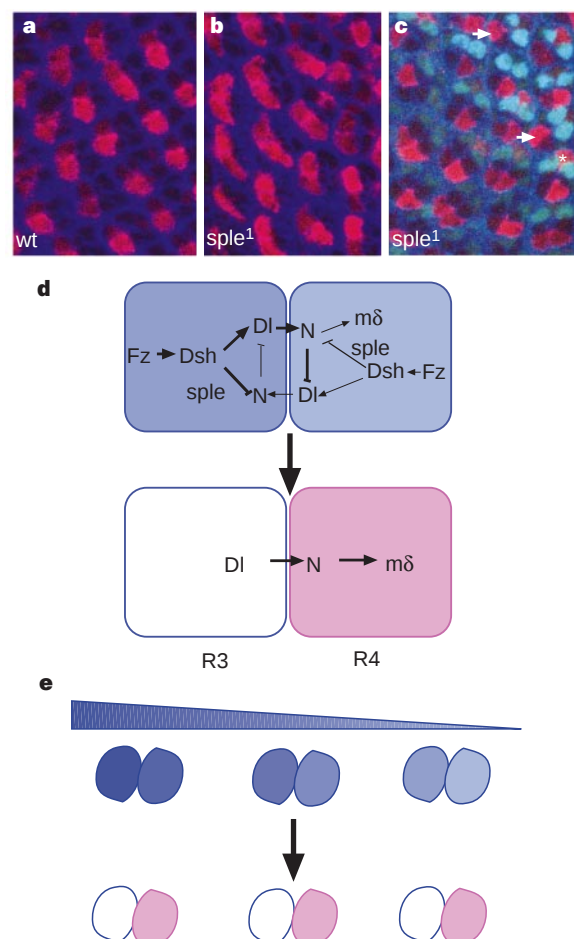


Figure 4. Notch signalling in R3/R4 suggests a mechanism to convert graded Fz/Dsh signalling into two distinct cell fates. **a-c**, Elevated mδ0.5 expression (red) in *sple¹*; rows 4-7 (**a**, wild type; **b**, *sple¹*) and 6-9 (**c**, *sple¹*); arrows indicate equatorial cell with mδ0.5 expression; asterisk indicates R7 with ectopic mδ0.5. Anti-coracle blue, anti-Bar turquoise (R1/R6). **d**, Scheme depicting interaction between Fz/Dsh and Dl/N pathways in R3/R4; font size indicates relative levels of activation and blue shading represents Fz activation high (dark) low (light). As a consequence of differential activation of Dl/N and subsequent feedback, the two cells acquire distinct identities, R3 (blue) and R4 (pink). **e**, Hypothetical gradient (blue) from the equator (dark) to the poles (light). Adjacent cells receive slightly different amounts. Through the scheme in **d** this would be converted to two distinct cell states, R3 and R4.

locus. In agreement, the wing phenotypes (thickened veins and wing-edge loss) of N^{md-3} mutants are completely suppressed by $pk-sple^{13}$ (data not shown). The inhibitory effect of Sple provides a second mechanism for polarizing Notch signalling in R3/R4 that could be coordinated by Fz/Dsh (Fig. 4d).

In conclusion, Notch signalling is activated in R3/R4 in response to Fz/Dsh and we propose that Fz/Dsh sets up an initial bias in Notch activity between R3 and R4 by promoting Df activity and inhibiting Notch via Sple in a coordinated manner. Feedback in the Notch pathway amplifies this bias so that even a subtle variation in the amount of signal received by Fz in the equatorial (pre-R3) and polar (pre-R4) cells in each ommatidium would generate distinct cell fates. This explains how a signal from the equator could be interpreted by the whole field of ommatidia (Fig. 4e) and is likely to be of widespread significance in the development of polarized structures within planar epithelia. The asymmetrical expression of Notch-pathway genes detected in feather primordia is consistent with this model²⁴. Furthermore, this mechanism for the coordinated regulation of Notch signalling can also explain how neural precursors develop at specific positions within competent proneural fields. □

Methods

Fly strains. Alleles used were: N^{ts1} , N^{md-3} , dsh^1 , fz^{R54} , Df^1 , Df^{6B} , $sple^1$ and $pk-sple^{13}$. For analysis of $E(spl)$ a combination between $Df(3R)/NFI^{P1}$ (removing NFI , $E(spl)md$ and $E(spl)mry$ promoter²⁵; M.T.D.C. and S.J.B., unpublished data) and $Df(3R)/E(spl)^{grob32.2}$ (removing all $E(spl)$ genes and NFI) was used. Rescue of NFI activity²⁵ did not modify the eye phenotypes and no defects were observed in $NFI^{P2}/Df(3R)/NFI^{P1}$, which eliminates NFI only²⁵. For mis-expression studies we used heat-shock-inducible, intracellular Notch ($hs-N^{icd}$; ref. 15), a transmembrane-activated Notch derivative driven by *sevenless* enhancer (sev^E-N^{icd} ; ref. 17), and the Gal4/UAS-targeted expression system²⁶. UAS lines were: $UAS-N^{icd}$ (gift of M. Haenlin), $UAS-N^{dm}$ (ECN, containing Notch extracellular and transmembrane domains¹⁶), $UAS-dsh^{19}$, $UAS-E(spl)md$ and $UAS-E(spl)mb$. These were combined with $sev-Gal4$ (expressed in R3, R4, R7, mystery and cone cells) and/or $spalt-Gal4$ (expressed in R3, R4 and cone cells).

For N^{ts} experiments, larvae were incubated at 30 °C for 6 h, returned to 25 °C for 10 h or until eclosion. N^{icd} expression was induced in $hs-N^{icd}$ larvae by 2 h at 37 °C.

md0.5 transgenic lines. A 487-bp fragment from the 1.9-kb genomic *HindIII* fragment upstream of $E(spl)md$ was amplified using the primers GATCTA-GATGCCATCAGATGTCAGC and CTACTAGTCTTTTGGCGCACAGTCAC, digested with *SpeI* (filled in) and *XbaI*, and ligated into *Asp718* (filled in) and *XbaI* sites of HZ50PL. Transgenic lines were established by injection in *cn;ry* flies using standard procedures and all lines gave identical patterns of expression.

Immunofluorescence. the following antibodies were used: rabbit anti- β -galactosidase (Cappell), rabbit anti-Bar²⁷, guinea-pig anti-Coracle²⁸, rat anti-Elav (Developmental Studies Hybridoma Bank), rat anti-Spalt (a gift of R. Barrio) and mouse monoclonal antibodies against β -galactosidase (Promega), $E(spl)$ proteins²⁹, Rough³⁰ and Delta²⁰.

Received 26 October; accepted 18 December 1998.

- Tomlinson, A. Cellular interactions in the developing eye. *Development* **104**, 183–193 (1988).
- Wolff, T. & Ready, D. F. in *The Development of Drosophila melanogaster* (eds Bate, M. & Martinez-Arias, A.) 1277–1325 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1993).
- Zheng, L., Zhang, J. & Carthew, R. W. *frizzled* regulates mirror-symmetric pattern formation in the *Drosophila* eye. *Development* **121**, 3045–3055 (1995).
- Fanto, M., Mayes, C. A. & Mlodzik, M. Linking cell-fate specification to planar polarity: determination of the R3/R4 photoreceptors is a prerequisite for the interpretation of the *Frizzled*. *Mech. Dev.* **74**, 51–58 (1998).
- Wolff, T. & Rubin, G. M. *strabismus*, a novel gene that regulates tissue polarity and cell fate decisions in *Drosophila*. *Development* **125**, 1149–1159 (1998).
- Theisen, H. et al. *dishevelled* is required during *wingless* signaling to establish both cell polarity and cell identity. *Development* **120**, 347–360 (1994).
- Krasnow, R. E., Wong, L. L. & Adler, P. N. *dishevelled* is a component of the *frizzled* signaling pathway in *Drosophila*. *Development* **121**, 4095–4102 (1995).
- Strutt, D. I., Weber, U. & Mlodzik, M. The role of RhoA in tissue polarity and *Frizzled* signalling. *Nature* **387**, 292–295 (1997).
- Boutros, M., Paricio, N., Strutt, D. I. & Mlodzik, M. *Dishevelled* activates JNK and discriminates between JNK pathways in planar polarity and *wingless* signaling. *Cell* **94**, 109–118 (1998).
- Wehrli, M. & Tomlinson, A. Independent regulation of anterior/posterior and equatorial/polar polarity in the *Drosophila* eye: evidence for the involvement of Wnt signaling in the equatorial/polar axis. *Development* **125**, 1421–1432 (1998).
- de Celis, J. F. et al. Functional relationships between *Notch*, *Su(H)* and the bHLH genes of the $E(spl)$ complex: the $E(spl)$ genes mediate only a subset of *Notch* activities during imaginal development. *Development* **122**, 2719–2728 (1996).

- Eastman, D. S. et al. Synergy between suppressor of Hairless and Notch in the regulation of *Enhancer of split mry* and *md* expression. *Mol. Cell. Biol.* **17**, 5620–5628 (1997).
- Cagan, R. L. & Ready, D. F. Notch is required for successive cell decisions in the developing *Drosophila* retina. *Genes Dev.* **3**, 1099–1112 (1989).
- Baker, N. E., Sung, Y. & Han, D. Evolution of proneural *atonal* expression during distinct regulatory phases in the developing *Drosophila* eye. *Curr. Biol.* **6**, 1290–1301 (1996).
- Struhl, G., Fitzgerald, K. & Greenwald, I. Intrinsic activity of the Lin-12 and Notch intracellular domains *in vivo*. *Cell* **74**, 331–345 (1993).
- Klein, T., Brennan, K. & Arias, A. M. An intrinsic dominant negative activity of serrate that is modulated during wing development in *Drosophila*. *Dev. Biol.* **189**, 123–134 (1997).
- Fortini, M. E., Rebay, L., Caron, L. A. & Artavanis-Tsakonas, S. An activated Notch receptor blocks cell-fate commitment in the developing *Drosophila* eye. *Nature* **365**, 555–557 (1993).
- Fischer-Vize, J. A., Vize, P. D. & Rubin, G. M. A unique mutation in the *Enhancer of split* gene complex affects the fates of the mystery cells in the developing *Drosophila* eye. *Development* **115**, 89–101 (1992).
- Axelrod, J. D., Matsuno, K., Artavanis-Tsakonas, S. & Perrimon, N. Interaction between *Wingless* and Notch signaling pathways mediated by *Dishevelled*. *Science* **271**, 1826–1832 (1996).
- Parks, A. L., Turner, F. R. & Muskavitch, M. A. T. Relationships between complex Delta expression and the specification of retinal cell fates during *Drosophila* eye development. *Mech. Dev.* **50**, 201–216 (1995).
- Greenwald, I. LIN-12/Notch signaling: lessons from worms and flies. *Genes Dev.* **12**, 1751–1762 (1998).
- Choi, K. W., Mozer, B. & Benzer, S. Independent determination of symmetry and polarity in the *Drosophila* eye. *Proc. Natl Acad. Sci. USA* **93**, 5737–5741 (1996).
- Gubb, D. Cellular polarity, mitotic synchrony and axes of symmetry during growth. Where does the information come from? *Int. J. Dev. Biol.* **42**, 369–377 (1998).
- Chen, C.-W. J., Jung, H.-S., Jiang, T.-X. & Chuong, C.-M. Asymmetric expression of Notch/Delta/Serrate is associated with the anterior-posterior axis of feather buds. *Dev. Biol.* **188**, 181–187 (1997).
- The, I. et al. Rescue of a *Drosophila* *NFI* mutant phenotype by Protein Kinase A. *Science* **276**, 791–794 (1997).
- Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
- Higashijima, S. et al. Dual *Bar* homeo box genes of *Drosophila* required in two photoreceptor cells, R1 and R6, and primary pigment cells for normal eye development. *Genes Dev.* **6**, 50–60 (1992).
- Fehon, R. G., Dawson, I. A. & Artavanis-Tsakonas, S. A *Drosophila* homolog of membrane skeleton protein-4.1 is associated with septate junctions and is encoded by the *coracle* gene. *Development* **120**, 545–547 (1994).
- Jennings, B., Preiss, A., Delidakis, C. & Bray, S. The Notch signalling pathway is required for *Enhancer of split* bHLH protein expression during neurogenesis in the *Drosophila* embryo. *Development* **120**, 3537–3548 (1994).
- Kimmel, B. E., Heberlein, U. & Rubin, G. M. The homeodomain protein Rough is expressed in a subset of cells in the developing *Drosophila* eye where it can specify photoreceptor subtype. *Genes Dev.* **4**, 712–727 (1990).

Acknowledgements. We thank R. Barrio, J. de Celis, M. Dominguez, D. Gubb and M. Haenlin for reagents; M. Dominguez for valuable advice; the multi-imaging center for technical assistance, and N. Brown, K. Moses and M. Freeman for comments on the manuscript. This research was supported by a Project Grant from the MRC and M.T.D.C. was funded by an MRC studentship.

Correspondence and requests for materials should be addressed to S.J.B. (sjb32@mole.bio.cam.ac.uk).

Cyclosporine induces cancer progression by a cell-autonomous mechanism

Minoru Hojo*†, Takashi Morimoto‡, Mary Maluccio*, Tomohiko Asano§, Kengo Morimoto†, Milagros Lagman*, Toshikazu Shimbo† & Manikkam Suthanthiran*

* Department of Transplantation Medicine and Extracorporeal Therapy, Division of Nephrology, and Departments of Medicine and Surgery, Weill Medical College of Cornell University, 525 East 68th Street, New York, New York 10021, USA
† Department of Pediatrics, Mizonokuchi Hospital, Teikyo University School of Medicine, 3-8-3 Mizonokuchi, Takatsu-ku, Kawasaki 213, Japan
‡ Department of Cell Biology, New York University School of Medicine, 550 First Avenue, New York, New York 10016, USA
§ Department of Urology, National Defense Medical College, 3-2 Namiki, Tokorozawa, Saitama 359, Japan

Malignancy is a common and dreaded complication following organ transplantation^{1–4}. The high incidence of neoplasm and its aggressive progression, which are associated with immunosuppressive therapy, are thought to be due to the resulting impairment of the organ recipient's immune-surveillance system^{5–9}. Here we report a mechanism for the heightened malignancy that is independent of host immunity. We show that cyclosporine (cyclosporin A), an immunosuppressant that has had a major impact on improving patient outcome following organ transplantation^{4,5}, induces phenotypic changes, including invasiveness of non-transformed cells, by a cell-autonomous mechanism.

Our studies show that cyclosporine treatment of adenocarcinoma cells results in striking morphological alterations, including membrane ruffling and numerous pseudopodial protrusions, increased cell motility, and anchorage-independent (invasive) growth. These changes are prevented by treatment with monoclonal antibodies directed at transforming growth factor- β (TGF- β). *In vivo*, cyclosporine enhances tumour growth in immunodeficient SCID-beige

mice; anti-TGF- β monoclonal antibodies but not control antibodies prevent the cyclosporine-induced increase in the number of metastases. Our findings suggest that immunosuppressants like cyclosporine can promote cancer progression by a direct cellular effect that is independent of its effect on the host's immune cells, and that cyclosporine-induced TGF- β production is involved in this.

We explored an alternative and autonomous cellular mechanism

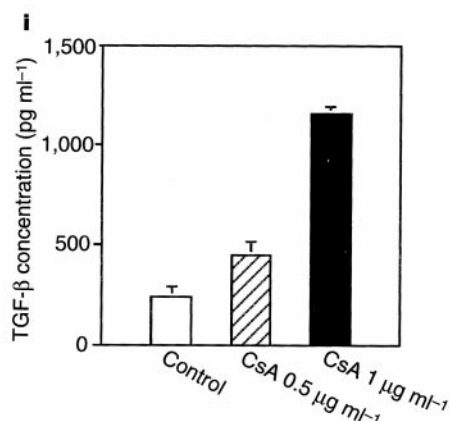
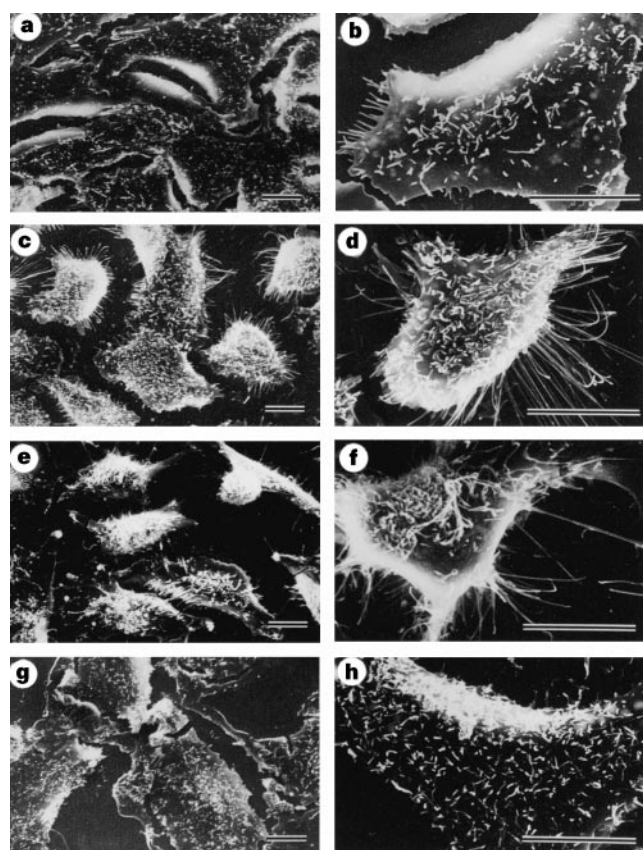


Figure 1 Cyclosporine induces A-549 cells to acquire an invasive phenotype. Scanning electron microscopic photographs of A-549 cells grown on glass coverslips and incubated for 72 h with: nothing (control cells, **a**, **b**); 1 μ g ml⁻¹ cyclosporine (**c**, **d**); 2 ng ml⁻¹ recombinant TGF- β ₁ protein (**e**, **f**); 1 μ g ml⁻¹ cyclosporine plus 30 μ g ml⁻¹ anti-TGF- β (**g**, **h**). Note that cyclosporine-conditioned cells, in a similar fashion to TGF- β ₁-treated cells, display membrane ruffling and acquire exploratory pseudopodia, and that anti-TGF- β antibodies prevent these cyclosporine-induced morphological alterations. Scale bars, 10 μ m. **i**, TGF- β concentrations (mean $\pm$ s.d.) in supernatants obtained from untreated or cyclosporine-treated A-549 cells. The cells were incubated for 72 h, in the absence or presence of 0.5 or 1.0 μ g ml⁻¹ cyclosporine, and a sandwich ELISA assay¹¹ was used to quantify TGF- β levels.

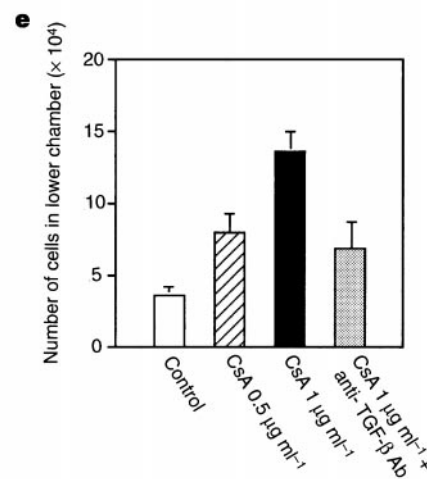
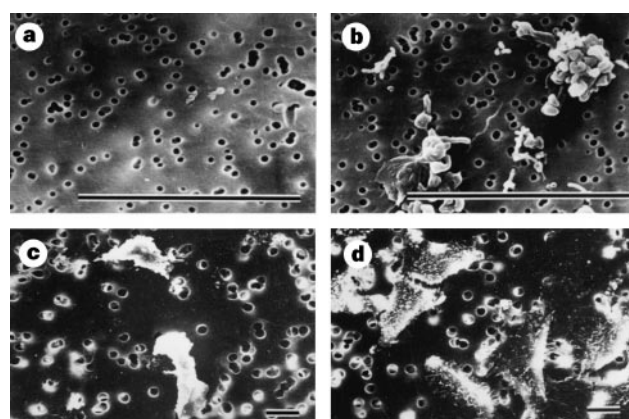


Figure 2 Cyclosporine stimulates the motility of A-549 cells. A-549 cells were grown for 72 h on 10-mm round polycarbonate membrane filters with 0.4 μ m (**a**, **b**) or 3 μ m (**c**, **d**) pore size in the absence (**a**, **c**) or presence (**b**, **d**) of 1 μ g ml⁻¹ cyclosporine. The filters were removed from the culture dishes and the bottom surfaces were examined by scanning electron microscopy. Note that the pseudopodia of cyclosporine-treated cells grown on 0.4- μ m pore size filters protrude through the pores to the bottom surface (**b**) whereas cyclosporine-treated cells grown on 3- μ m pore filters migrate through the pores onto the bottom surface (**d**). Scale bars, 10 μ m. **e**, Cyclosporine promotes migration of A-549 cells. A-549 cells were placed in the upper chamber of 12-well transwells (8 μ m pore) at a density of 10^5 cells, and were incubated alone or with 0.5 or 1.0 μ g ml⁻¹ cyclosporine or with 1.0 μ g ml⁻¹ cyclosporine plus 30 μ g ml⁻¹ anti-TGF- β antibodies (Ab). The cells that migrated into the lower chamber through the 8- μ m pores of the polycarbonate membrane filters were counted after the cells had been dissociated by trypsinization. Results (mean $\pm$ s.e.m.) are of three experiments carried out with duplicate samples.

for immunosuppressant-associated tumour progression. We tested the hypothesis that cyclosporine, independently of any effects on the host immune system, would programme non-invasive cells to acquire an invasive phenotype. The experimental basis for our hypothesis was our demonstration that cyclosporine promotes the transcription and functional expression of the TGF- β_1 gene^{10,11} and the observation by others that TGF- β can promote tumour-cell invasion and metastasis¹²⁻¹⁴.

A non-transformed human pulmonary adenocarcinoma (A-549) cell line¹⁴ that is not invasive *in vitro* was used as the indicator cell to test the hypothesis that cyclosporine can induce an invasive phenotype. A-549 cells express functional receptors for TGF- β , and their growth and function are regulated by TGF- β (refs 11, 14). We carried out these experiments *ex vivo* to avoid any confounding effects of cyclosporine-associated inhibition of *in vivo* immune surveillance mechanisms.

Figure 1 shows the striking morphological changes observed following cyclosporine treatment of A-549 cells. Scanning electron microscopic examination revealed that untreated A-549 cells display a cuboidal epithelial and non-invasive phenotype (Fig. 1a, b), whereas cyclosporine-treated cells show phenotypic alterations that are characteristic of invasive cells, that is, marked membrane ruffling and the formation of numerous pseudopodia (Fig. 1c, d). Additional data (Fig. 1) support the hypothesis that the cyclosporine-induced acquisition of an invasive phenotype is due

to TGF- β . First, cyclosporine stimulated TGF- β secretion by A-549 cells in a concentration-dependent manner (Fig. 1i); second, anti-TGF- β monoclonal antibodies (1D11.16 IgG)₁¹⁵, in contrast to control IgG₁ monoclonal antibodies, prevented the cyclosporine-induced morphological alterations (Fig. 1g, h); and third, recombinant TGF- β_1 induced morphological alterations in A-549 cells that were similar to those elicited by cyclosporine (Fig. 1e, f). Our finding that cyclosporine stimulates TGF- β_1 production in A-549 cells extends earlier observations that it induces T cells¹⁰, CCL-64 mink lung epithelial cells¹¹ and renal cells¹⁶ to hyperexpress TGF- β_1 . The phenotypic changes elicited by cyclosporine were reversible; incubation of cyclosporine-treated A-549 cells in cyclosporine-free culture medium for 48 h resulted in the reversal of the invasive phenotype and a return to the original morphology (data not shown).

Cells capable of locomotion and invasiveness display exploratory pseudopodia^{17,18}. Because cyclosporine induced numerous, long pseudopodia in A-549 cells, we investigated whether such cells acquired motility. To explore this, A-549 cells were seeded on polycarbonate membrane filters with three pore sizes (0.4, 3 and 8 μ m) in the presence or absence of cyclosporine; and the bottom surfaces of the membrane filters were examined by scanning electron microscopy. Our results show that many cyclosporine-induced pseudopodia protrude through 0.4- μ m pores onto the bottom surface (Fig. 2b), whereas only a few pseudopodia protrude in the control (Fig. 2a). When the cells were grown on 3- μ m pore filters, many cyclosporine-treated cells and only few untreated cells, migrated through the pores to the bottom surface of the membrane filter (Fig. 2c, d).

To quantify the cell motility resulting from cyclosporine treatment, we used 8- μ m pore filters in the migration assay. We found that the number of A-549 cells that migrated increased in proportion to the concentration of cyclosporine used to treat the cells, and that the increased cell motility was suppressed by anti-TGF- β antibodies (Fig. 2e), but not by the control IgG₁ antibodies. Thus, cyclosporine-induced alterations in both morphology (Fig. 1) and cell motility (Fig. 2) were dependent on cyclosporine-induced TGF- β production.

Anchorage-independent growth *in vitro* is considered a correlate of invasive tumour growth *in vivo*^{19,20}, and so we next examined whether cyclosporine treatment results in anchorage-independent growth. A-549 cells were plated on soft agarose gels and grown for 96 h (Fig. 3, see legend for details). Phase-contrast microscopic examination revealed that untreated A-549 cells retained their spherical shape and remained suspended in the culture medium, whereas cyclosporine-treated cells spread and proliferated strongly on the soft gel (Fig. 3a, b). Because it was difficult to determine by phase-contrast microscopy whether the pseudopodia extended along the surface of the agarose or penetrated deeper into the agarose layer, we made vertical thin sections of the soft gels and examined them by scanning electron microscopy to obtain side views. This strategy revealed that many fully grown pseudopodia of the cyclosporine-treated cells penetrated the agarose-gel layer and extended vertically into the gel plate (Fig. 3d). Also, these cells appeared to be supported by the extensively invaded pseudopodia, in contrast to the absence of pseudopodial extensions in the control A-549 cells (Fig. 3c).

Cyclosporine's effect on A-549 cells growth was contingent upon whether the culture conditions were anchorage-dependent or -independent. It inhibited the proliferation of A-549 cells under anchorage-dependent conditions but stimulated proliferation under anchorage-independent conditions (Fig. 3e, f).

We next examined whether cyclosporine induces morphological and functional alterations in other cell types, looking at murine renal cell adenocarcinoma (Renca) cells, mouse mammary gland epithelial (NMuMG) cells and mink lung epithelial (CCL-64) cells. We found that cyclosporine treatment produced phenotypic

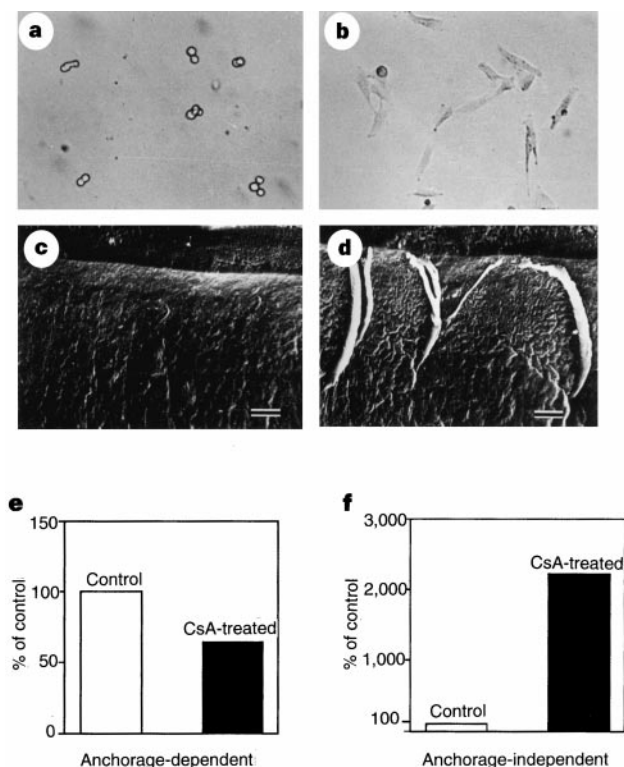


Figure 3 Anchorage-independent growth of cyclosporine-conditioned A-549 cells. Culture medium (3 ml) containing 10^4 cells and $1 \mu\text{g ml}^{-1}$ cyclosporine (**b, d**) or no cyclosporine (**a, c**) was loaded onto 5 ml of a 0.3% agarose layer containing MEM-10% FBS in 60-mm dishes. After 96 h incubation at 37 °C, cells grown on the surface of the agarose layer were examined with a phase-contrast microscope (**a, b**). For scanning electron microscopic observation (**c, d**), the soft gels were fixed and then vertical thin sections were made. These slices were processed for scanning electron microscopy as described in Methods. Twenty slices were made from each agarose plate, and representative ones are shown. Note that pseudopodia of cyclosporine-treated A-549 cells protrude deeply into the soft gel under anchorage-independent growth conditions. Scale bars, 10 μ m. **e**, Cyclosporine-associated inhibition of anchorage-dependent proliferation of A-549 cells; **f**, cyclosporine-induced stimulation of anchorage-independent growth.

alterations in these epithelial cells as well. A representative example (Fig. 4a, b) shows that cyclosporine-treated Renca cells, in a similar fashion to cyclosporine-treated A-549 cells, display an invasive phenotype.

We investigated whether cyclosporine would enhance the invasive and metastatic growth of tumour cells *in vivo* using Renca cells²¹, and two other tumour cell lines, mouse-derived Lewis lung carcinoma cells²² and human bladder transitional carcinoma cells²³, as the tumour inoculum. SCID-beige mice (mice homozygous for both *SCID* and *beige* mutations²⁴), which are deficient in T cells, B cells and natural killer cells, were used as the host. The use of SCID-beige mice minimized the possibility that cyclosporine-induced depression of the host's immune system contributed to tumour progression.

Cyclosporine increased the number of murine renal carcinoma metastases in SCID-beige mice (Fig. 4c, d). Data from four separate experiments showed that the number of renal cell cancer pulmonary metastases was 241 ± 22 (mean $\pm$ s.e.m., $n = 21$) in the control SCID-beige mice compared with 338 ± 26 ($n = 18$) in the cyclosporine-treated mice ($P = 0.007$; *t*-test) (Table 1). Also, the number of pulmonary metastases resulting from inoculation of murine Lewis lung carcinoma cells was 11 ± 2 ($n = 9$ mice) in the control mice compared with 28 ± 4 ($n = 8$ mice) with cyclosporine treatment ($P = 0.003$), while the number of pulmonary metastases resulting from inoculation of human bladder transitional cancer cells was 63 ± 18 ($n = 9$ mice) in controls and 138 ± 21 ($n = 9$ mice) in cyclosporine-treated mice ($P = 0.01$) (Table 1).

We investigated the effect of anti-TGF- β antibodies (1D11.16 IgG₁)¹⁵ on the cyclosporine-induced increase in the metastases to determine whether *in vivo* tumour progression by cyclosporine was dependent on TGF- β . Anti-TGF- β antibodies, but not control IgG₁ antibodies, prevented the cyclosporine-induced increase in metastases. The number of pulmonary metastases was 350 ± 22 (mean $\pm$ s.e.m., $n = 12$) in the control mice, 441 ± 20 ($n = 10$) in cyclosporine-treated mice, 284 ± 34 ($n = 8$) in mice treated with both cyclosporine and anti-TGF- β , and 490 ± 56 ($n = 4$) in mice treated with cyclosporine and control IgG₁ ($P = 0.0005$; one-way ANOVA). The reduction in the number

Table 1 Cyclosporine increases pulmonary metastases in SCID-beige mice

Tumour inoculum	Number of pulmonary metastases (mean $\pm$ s.e.m.)		<i>P</i> *
	Without CsA	With CsA	
Murine Renca	241 ± 22 ($n = 21$)	338 ± 26 ($n = 18$)	0.007
Murine Lewis lung carcinoma (LLC)	11 ± 2 ($n = 9$)	28 ± 4 ($n = 8$)	0.003
Human bladder cancer (T24)	63 ± 18 ($n = 9$)	138 ± 21 ($n = 9$)	0.01

Tumour cells (1×10^6 or 5×10^5 in HBSS) were injected in the tail vein of SCID-beige mice. Cyclosporine (cyclosporin A; CsA; 20 mg per kg) was administered every other day from day -1 to the day of death. The mice were killed on day 19 (Renca), 16 (LLC) or 23 (T24), and the number of metastases was counted as described³⁰. *n*, Number of mice in each group; *P* value derived with *t*-test.

of metastases found following the administration of anti-TGF- β antibodies to cyclosporine-treated mice was significant at $P < 0.01$ by ANOVA (Bonferroni *P*-value). In contrast, there was no significant difference between the number of metastases found in cyclosporine-treated mice and that found in mice treated with combined cyclosporine and control IgG₁ ($P > 0.05$). Our *in vitro* experiments show that the tumour cells are the sole source of TGF- β (Figs 1, 2). Many cell types, in addition to tumour cells, might contribute to cyclosporine-induced TGF- β hyperexpression *in vivo*.

The malignancy-promoting effects of immunosuppressive drugs are thought to result from drug-induced T-lymphocyte dysfunction and resultant immunosuppression. On the other hand, the production of TGF- β by tumours represents a potential mechanism by which they evade the host's immune system²⁵⁻²⁸. Our demonstration that cyclosporine-treated, non-transformed cells acquire invasiveness under *in vitro* conditions that allow no possible involvement of the host's immune system, and our *in vivo* data that cyclosporine promotes tumour growth in SCID-beige mice, suggest a cell-autonomous mechanism for cancer progression (Fig. 5). Specific therapeutic strategies that target pathways responsible for heightened invasiveness (such as TGF- β regulation) are worth exploring and may be of value to people who are given allografts and to other individuals at increased risk of neoplasms. □

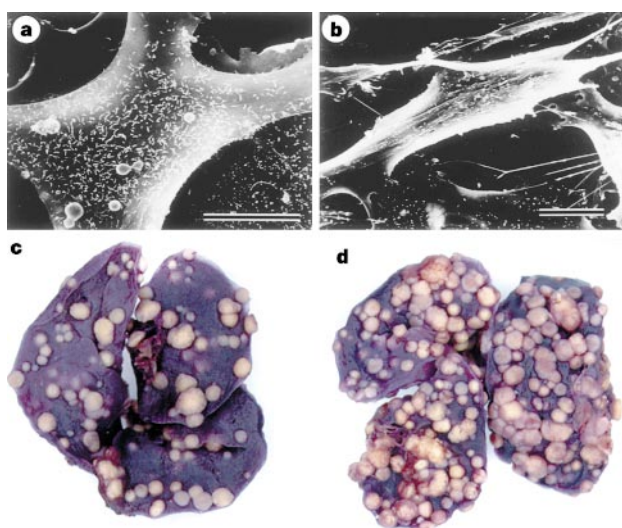


Figure 4 Cyclosporine induces renal cancer cells to acquire an invasive phenotype and promotes tumour growth *in vivo*. Scanning electron micrograph of murine renal adenocarcinoma cells incubated for 72 h in the absence (a) or presence of $1 \mu\text{g ml}^{-1}$ cyclosporine (b). Scale bars, $10 \mu\text{m}$. Representative lungs, retrieved from untreated mice (c) and from cyclosporine-treated mice (d), are shown to illustrate the cyclosporine-associated increase in renal cell cancer pulmonary metastasis in SCID-beige mice.

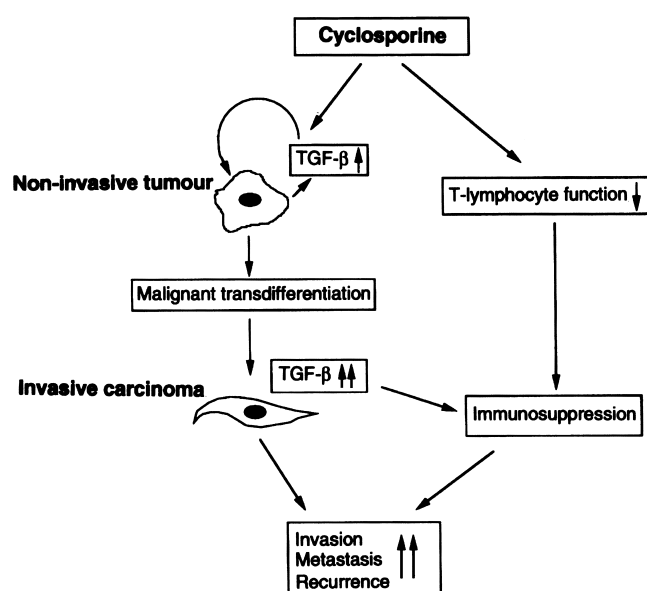


Figure 5 Potential mechanisms for cyclosporine-associated tumour progression. In this formulation, cyclosporine induced TGF- β production by tumour cells promotes cell invasiveness by a cell-autonomous mechanism that is independent of and/or complementary to cyclosporine's immunosuppressive effect on the host's immune system.

Methods

Cell line and culture. Human lung adenocarcinoma cells (A-549 cells; ATCC CCL 185; American Type Culture Collection, Rockville, MD), human bladder transitional carcinoma cells (ATCC HTB4, T24), mink lung epithelial cells (CCL-64; ATCC), mouse mammary gland epithelial cells (NMuMG; ATCC) and Lewis lung carcinoma cells (ATCC) were grown in minimum essential medium (MEM) supplemented with 10% fetal bovine serum (FBS), at 37°C in a 95% air–5% CO₂ atmosphere. Murine renal adenocarcinoma cells (Renca cells; a gift from R. H. Wiltout, National Cancer Institute) were maintained by *in vivo* serial passages in syngeneic BALB/c mice, as described²¹.

Scanning electron microscopy. Cells were seeded at a density of 10⁵ on 12-mm round glass coverslips or 10-mm round polycarbonate membrane filters (0.4 or 3 µm pore size) in 12-well Transwells (Costar), and grown for 72 h in the presence or absence of cyclosporine. To assess the ability of TGF-β-specific antibody to inhibit cyclosporine-mediated effects, A-549 cells were incubated in the presence of both cyclosporine and monoclonal antibodies (Genzyme) that recognize TGF-β₁, -β₂ and -β₃. Cells were fixed with PBS, pH 7.4, containing 2.0% glutaraldehyde, and processed as previously described²⁹. Samples were examined using a JEOL 25SIII electron microscope.

Quantification of TGF-β. TGF-β was quantified using a sandwich enzyme-linked immunosorbent assay (ELISA) method as previously described¹¹. In brief, each well of multiwell ELISA assay plates was coated with anti-TGF-β₁ antibodies (1 µg ml⁻¹). The plates were incubated for 2 h at 37°C after the addition of various amounts of TGF-β₁ in PBS or conditioned medium. After washing with PBS containing 0.2% Tween-20 (PBST), rabbit antiserum against TGF-β was added to each well. The plates were incubated at 37°C for 1 h, the wells were washed with PBST, and then 100 µl of goat anti-rabbit IgG–alkaline phosphatase conjugates was added. Absorbance at 430 nm was measured using an ELISA assay reader. A-549 cells were cultured in serum-free medium to exclude contamination of cell-free supernatants by serum-derived TGF-β.

Cell proliferation assay. For assaying anchorage-dependent growth, A-549 cells were grown at a density of 2 × 10⁴ cells per well in 12-well plates in the presence or absence of cyclosporine. After 96 h treatment, each well received 2 µCi of methyl-³H-thymidine, and cells were incubated for an additional 4 h. They were washed twice with ice-cold PBS and fixed with methanol for 60 min. After washing, the fixed cells were lysed with 0.2 M NaOH and treated with cold 10% trichloroacetic acid (TCA) for 15–20 min on ice. The radioactivity, recovered as cold TCA-insoluble precipitates, was used for measuring relative cell proliferation by comparing the radioactivity between control and experiment. For an anchorage-independent cell growth, cells spread well on agarose gel were counted using a phase-contrast microscope.

***In vivo* tumour growth.** Murine renal cell adenocarcinoma cells (1 × 10⁵ in Hank's balanced salt solution; HBSS), murine Lewis lung carcinoma cells (5 × 10⁵ cells) or human bladder cancer cells (1 × 10⁵ cells) were injected in the tail vein of 6-week-old male SCID–beige mice. Cyclosporine (20 mg per kg in 0.2 ml olive oil) was administered every other day starting from day –1, to day 19 or 23 after tumour inoculation. On day 19 or 23 after tumour inoculation, mice were killed and the number of pulmonary metastases was determined³⁰ following endotracheal insufflation of lungs with 15% India ink solution and bleaching the collected lungs in Fekete's solution. The effect of anti-TGF-β antibody and the control IgG₁ antibody on the cyclosporine-induced increase in the number of pulmonary metastases was determined by intraperitoneal administration of 200 µg of antibody, on a daily basis starting from day –1 to day 19 after tumour inoculation.

Received 21 September; accepted 15 December 1998.

- Penn, I. Cancers following cyclosporine therapy. *Transplantation* **43**, 32–35 (1987).
- Yokoyama, I., Carr, B., Saito, H., Iwatsuki, S. & Starzl, T. E. Accelerated growth rates of recurrent hepatocellular carcinoma after liver transplantation. *Cancer* **68**, 2095–2100 (1991).
- London, N. J., Farmery, S. M., Will, E. J., Davison, A. M. & Lodge, J. P. A. Risk of neoplasia in renal transplant patients. *Lancet* **346**, 403–406 (1995).
- Suthanthiran, M. & Strom, T. B. Renal transplantation. *N. Engl. J. Med.* **331**, 365–376 (1994).
- Kahan, B. D. Cyclosporine. *N. Engl. J. Med.* **321**, 1725–1738 (1989).
- Shimizu, T., Martin, M. S., Pelletier, H., Lagadee, P. & Martin, F. Effects of cyclosporin A on progressive and regressive tumors induced by two cancer lines derived from a single colon carcinoma chemically induced in the rat. *Immunobiology* **178**, 401–415 (1989).
- Van de Vrie, W., Marquet, R. L. & Eggermont, A. M. M. Cyclosporin A enhances locoregional metastasis of the CC531 rat colon tumor. *J. Cancer Res. Clin. Oncol.* **123**, 21–24 (1997).
- Van der Elst, J. et al. Quantitative study of liver metastasis from colon cancer in rats after treatment with Cyclosporine A. *J. Natl Cancer Inst.* **77**, 227–232 (1986).
- Masuhara, M., Ogasawara, H., Katyal, S. L., Nakamura, T. & Shinozuka, H. Cyclosporine stimulates

hepatocyte proliferation and accelerates development of hepatocellular carcinomas in rats. *Carcinogenesis* **14**, 1579–1584 (1993).

- Li, B. et al. Differential regulation of transforming growth factor-β and interleukin 2 genes in human T cells: Demonstration by usage of novel competitor DNA constructs in quantitative polymerase chain reaction. *J. Exp. Med.* **174**, 1259–1262 (1991).
- Khanna, A., Li, B., Stenzel, K. H. & Suthanthiran, M. Regulation of new DNA synthesis in mammalian cells by cyclosporine. *Transplantation* **57**, 577–582 (1994).
- Welch, D. R., Fabra, A. & Nakajima, M. Transforming growth factor-β stimulates mammary adenocarcinoma cell invasion and metastatic potential. *Proc. Natl Acad. Sci. USA* **87**, 7678–7682 (1990).
- Caulin, C., Scholl, F. G., Frontelo, P., Gamallo, C. & Quntanilla, M. Chronic exposure of cultured transformed mouse epidermal cells to transforming growth factor-β₁ induces an epithelial–mesenchymal transdifferentiation and a spindle tumoral phenotype. *Cell Growth Differ.* **6**, 1027–1035 (1995).
- Mooradian, D. L., McCarthy, J. B., Komanduri, K. V. & Furcht, L. T. Effects of transforming growth factor-β₁ on human pulmonary adenocarcinoma cell adhesion, motility, and invasion *in vitro*. *J. Natl Cancer Inst.* **84**, 523–527 (1992).
- Dasch, J. R., Pace, D. R., Waigell, W., Inenaga, D. & Ellingsworth, L. Monoclonal antibodies recognizing transforming growth factor-β. Bioactivity neutralization and transforming growth factor-β₂ affinity purification. *J. Immunol.* **142**, 1536–1541 (1989).
- Wolf, G., Thais, F. & Stahl, R. A. Cyclosporine stimulates expression of transforming growth factor-β in renal cells: Possible mechanism of cyclosporine's anti-proliferative effects. *Transplantation* **60**, 237–241 (1995).
- Guirguis, R., Margulies, I., Taraboletti, G., Schiffmann, E. & Liotta, L. Cytokine-induced pseudopod protrusion is coupled to tumor cell migration. *Nature* **329**, 261–263 (1987).
- Couchman, J. R., Lenn, M. & Rees, D. A. Coupling of cytoskeletal functions for fibroblast locomotion. *Eur. J. Cell Biol.* **36**, 182–194 (1985).
- Freeman, V. H. & Shin, S. Cellular tumorigenicity in nude mice: Correlation with cell growth in semi-solid medium. *Cell* **3**, 355–359 (1974).
- Massagué, J. Transforming growth factor-β family. *Annu. Rev. Cell Biol.* **6**, 597–641 (1990).
- Murphy, G. P. & Hruschsky, W. J. A murine renal cell carcinoma. *J. Natl Cancer Inst.* **50**, 1013–1025 (1973).
- Bertram, J. S. & Janik, P. Establishment of a cloned line of Lewis Lung Carcinoma cells adapted to cell culture. *Cancer Lett.* **11**, 63–73 (1980).
- Bubenik, J. et al. Established cell line of urinary bladder carcinoma (T24) containing tumour-specific antigen. *Int. J. Cancer* **11**, 765–773 (1973).
- Mosier, D. E., Stell, K. L., Gulizia, R. J., Torbett, B. E. & Gilmore, G. L. Homozygous scid/scid:beige/beige mice have low levels of spontaneous or neonatal T cell-induced B cell generation. *J. Exp. Med.* **177**, 191–194 (1993).
- Cui, W. et al. TGF-β₁ inhibits the formation of benign skin tumors, but enhances progression to invasive spindle carcinomas in transgenic mice. *Cell* **86**, 531–542 (1996).
- Atreaga, C. I. et al. Anti-transforming growth factor (TGF)-β antibodies inhibit breast cancer cell tumorigenicity and increase mouse spleen natural killer cell activity. *J. Clin. Invest.* **92**, 2569–2576 (1993).
- Torre-Aminoe, G. et al. A highly immunogenic tumor transfected with a murine transforming growth factor-β₁ cDNA escapes immune surveillance. *Proc. Natl Acad. Sci. USA* **87**, 1486–1490 (1990).
- Arrick, B. A. et al. Altered metabolic and adhesive properties and increased tumorigenesis associated with increased expression of transforming growth factor-β₁. *J. Cell Biol.* **118**, 715–726 (1992).
- Cohen-Gould, L., Robinson, T. F. & Factor, S. M. Intrinsic connective tissue abnormalities in the heart muscle of cardiomyopathy Syrian hamsters. *Am. J. Pathol.* **127**, 327–334 (1987).
- Asano, T., Khanna, A., Lagman, M., Li, B. & Suthanthiran, M. Immunostimulatory therapy with anti-CD3 monoclonal antibodies and recombinant interleukin-2: Heightened *in vivo* expression of mRNA encoding cytotoxic attack molecules and immunoregulatory cytokines and regression of murine renal cell carcinoma. *J. Urol.* **157**, 2396–2401 (1997).

Acknowledgements. We thank Y. Suzuki for technical assistance, B. Li, R. Ding, A. Khanna, V. K. Sharma and A. Niwa for their help; P. August, M. Ono and Y. Ichiyoshi for critical review of the manuscript; B. Pratt, D. Pappas and D. Felson for the generous gift and advice regarding 1D11.16 antibodies direct at TGF-β; and L. Stackhouse for help in the preparation of this manuscript. This work was supported in part by a grant from the Ministry of Education, Science and Culture, Japan (M.H.), Kanagawa Academy of Science and Technology, Japan (M.H.), and by an award from the National Institutes of Health, USA (M.S.).

Correspondence and requests for materials should be addressed to M.H.

RGD peptides induce apoptosis by direct caspase-3 activation

Christopher D. Buckley[†], Darrell Pilling[†], Nick V. Henriquez[†], Greg Parsonage^{*}, Katy Threlfall^{*}, Dagmar Scheel-Toellner^{*}, David L. Simmons^{*}, Arne N. Akbar[†], Janet M. Lord^{*} & Mike Salmon^{*}

^{*} Division of Immunity and Infection, and [†] Institute for Cancer Studies, MRC Centre for Immune Regulation, The University of Birmingham, Birmingham B15 2TT, UK

[§] Department of Neuroscience, SmithKline Beecham Pharmaceuticals, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 5AW, UK

[¶] Department of Clinical Immunology, Royal Free and University College Medical School, Pond Street, London NW3 2PF, UK

[†] These authors contributed equally to this work

Synthetic peptides containing the arginine–glycine–aspartate (RGD) motif have been used extensively as inhibitors of integrin–ligand interactions in studies of cell adhesion, migration, growth and differentiation^{1–3}, because the RGD motif is an integrin-recognition motif found in many ligands. Here we report

that RGD-containing peptides are able to directly induce apoptosis without any requirement for integrin-mediated cell clustering or signals. We show that RGD-containing peptides enter cells and directly induce autoprocessing and enzymatic activity of pro-caspase-3, a pro-apoptotic protein. Using the breast carcinoma cell line MCF-7, which has a functional deletion of the caspase-3 gene, we confirm that caspase-3 is required for RGD-mediated cell death. In addition to an RGD motif, pro-caspase-3 also contains a potential RGD-binding motif, aspartate-aspartate-methionine (DDM)⁴, near the site of processing to produce the p12 and p17 subunits⁵. On the basis of the ability of RGD-DDX interactions to trigger integrin activation⁶, we suggest that RGD peptides induce apoptosis by triggering conformational changes that promote pro-caspase-3 autoprocessing and activation. These findings provide an alternative molecular explanation for the potent pro-apoptotic properties of RGD peptides in models of angiogenesis, inflammation and cancer metastasis⁷⁻⁹.

During studies aimed at identifying factors involved in fibroblast-mediated T-cell survival¹⁰, we observed that peptides containing an RGD motif were able to induce apoptosis in resting (G0) peripheral blood T cells (Fig. 1a). Cell death was rapid (50% cell death occurring in 24 hours), dose-dependent and occurred over a low concentration range (0.1–1 mM RGD peptide), with half-maximal effect occurring at 250 μ M for the most potent peptide, GRGDNP. Only RGD-containing peptides induced apoptosis; peptides with conservative substitutions of single amino acids of identical charge (glutamine for aspartate or alanine for glycine) and a panel of

peptides containing random amino-acid sequences were ineffective. The minimal sequence required to induce apoptosis was the RGD motif itself (Fig. 1a). We confirmed that apoptosis was the route of death by using a range of standard methods¹⁰. The effect of these peptides on cell survival was not confined to resting peripheral blood T cells, as they also induced rapid apoptosis in CD4-positive T-cell lines, leukaemic T-cell lines (Jurkat and Molt-4 cells), B cells transformed by Epstein-Barr virus (LCL cells) and the erythroleukaemic cell line K562 (Fig. 1b). Thus the pro-apoptotic effects of RGD-containing peptides may be a general phenomenon occurring in resting, activated and transformed leukocytes. Two known T-cell

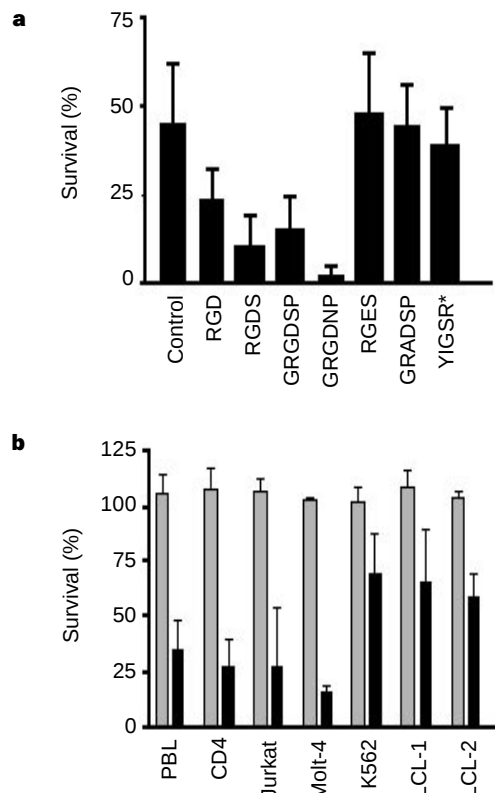


Figure 1 RGD-containing peptides induce lymphocyte apoptosis. **a**, Resting peripheral blood T cells were cultured for 3 days with peptides at concentrations of 1 mM (RGD, GRGDSP, GRGDNP, GRADSP and YIGSR) or 5 mM (RGDS and RGES) in serum-free media. Results are expressed as percentage survival of input cells. The YIGSR* peptide represents one of many control peptides that had no effect on survival. Results are shown as mean $\pm$ s.d., $n = 5$. **b**, A range of leukocyte cell types was cultured for 18 h in 1 mM GRGDNP (black bar) or GRADSP (grey bar). Survival values are shown as percentage survival compared with control (no peptide added). LCL-1 and LCL-2 represent two independent EBV-transformed B cell lines (mean $\pm$ s.d., $n = 5$). PBL, peripheral blood lymphocyte.

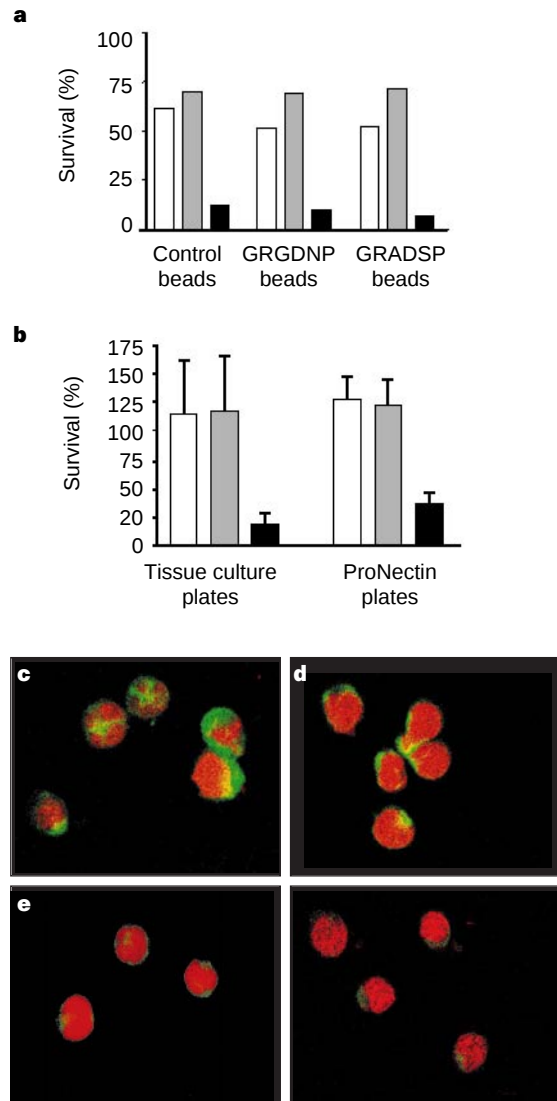


Figure 2 RGD peptides mediate their pro-apoptotic effects independently of integrin clustering and enter cells. **a**, Resting PBLs were cultured for 18 h with beads (ratio of 1:5 cells:beads) coupled with bovine serum albumin (control), GRGDNP or GRADSP. No peptide (open bar), free GRADSP (grey bar) or free GRGDNP (black bar) was added at 1 mM concentration. Data from one representative experiment of three are shown. Results are expressed as percentage survival of input cells. **b**, Molt-4 cells were cultured for 18 h in tissue culture plates or on ProNectin-F-coated plastic. No peptide (open bar), free GRADSP (grey bar) or free GRGDNP (black bar) was added at 1 mM concentration. Results are expressed as percentage survival of input cells (means $\pm$ s.d., $n = 3$). **c-f**, Molt-4 cells were cultured for 18 h with biotinylated GRGDNP (**c, e**) or GRADSP (**d, f**) before cells were fixed and stained (surface (red) and cytoplasmic (green) staining, **c, d**) or stained and then fixed (surface only staining, red, **e, f**) with streptavidin-FITC. Propidium iodide is the counter stain.

survival factors, interleukin (IL)-2 and fibroblast conditioned medium (IFN- β), which signal through distinct pathways¹⁰, were both unable to prevent RGD-mediated death.

As lymphocytes exhibit anchorage-independent growth and do not appear to require integrin-mediated signals to survive², we were surprised at the potent pro-apoptotic effects of RGD peptides in these cells. A possible explanation for the effects of these peptides was that they interfered with cell-cell interactions that might be required for leukocyte survival. We therefore studied the effect of cell density on the ability of RGD-containing peptides to induce cell death. When cultured at low cell density, which eliminated cell-cell interactions as observed microscopically, the RGD peptides still induced apoptosis of lymphocytes. Culturing the cells at low density effectively excluded any effects that these peptides might have on integrin-induced cell aggregation. As many integrins use an RGD motif to bind specific ligands, we next examined the effects on cell survival of a large panel of functional antibodies that recognize integrins known to be expressed on T cells. None of these antibodies (against integrins $\alpha 2$, $\alpha 3$, $\alpha 4$, $\alpha 5$, αV , αL , $\beta 1$, $\beta 2$ and $\beta 3$) were able to either support T-cell survival or induce T-cell death. These findings are fully consistent with our previous observations that, despite containing RGD motifs, matrix proteins such as fibronectin have no effect on T-cell survival¹⁰.

These data indicate that RGD-containing peptides may induce apoptosis directly by providing an active 'death' signal to leukocytes rather than inhibiting integrin-ligand-mediated survival signals. To address this possibility, we studied the pro-apoptotic effects of RGD-containing peptides that were bound to beads or covalently coupled to plastic as an array of repeating peptide motifs (ProNectin-F)¹¹. Multimeric presentation of ligand is known to favour receptor clustering and signalling¹². We therefore proposed that, when presented in this format, these peptides would induce more efficient apoptosis than soluble RGD peptides. Unexpectedly, when bound either to microbeads or to plastic surfaces, the pro-apoptotic effect of the RGD peptides was lost (Fig. 2a, b). However, the addition of soluble RGD-containing peptides restored apoptosis. The coating density of both RGD and control RAD peptides on the streptavidin beads was equal and saturating, as assessed by a competition assay using a biotinylated mouse antibody. Furthermore, the RGD-coated beads were functional, as assessed by their ability to induce cleavage of caspase-3 in a cell-free system (see below; Fig. 3c).

We next determined, by using labelled peptides and confocal microscopy, whether RGD- and RAD-containing peptides could be detected within cells. Both peptides were detected within the cytoplasm of treated cells (Fig. 2c-f). However, only RGD-containing

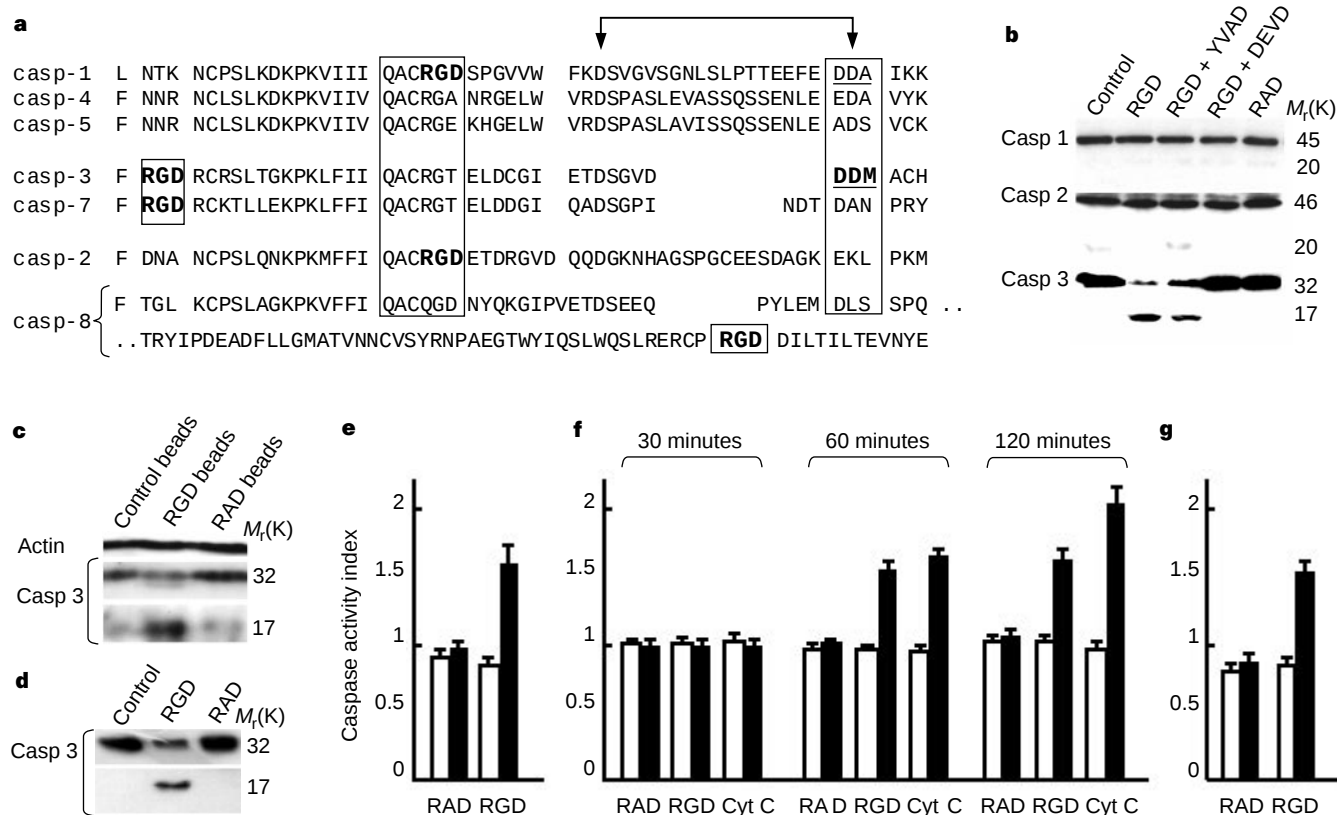


Figure 3 RGD-containing peptides induce pro-caspase-3 autoprocessing and activation directly in a cell-free system. **a**, Caspases 1 and 2 contain RGD sequences (beginning at R286 and R304, respectively) within the catalytically active site. Caspases 3 and 7 have homologous linear RGD motifs within the large subunit and caspase-8 has such a motif in the small subunit (motifs beginning at R144, R167 and R435, respectively). These sites are shown in bold within boxes. Potential RGD-binding sites D315DA and D180DM (in bold) near the known processing sites (arrows) for caspases 1 and 3 are underlined and boxed. **b**, Cell extracts from Molt-4 cells were incubated without peptide (control) or in the presence of 1 mM GRGDNP (RGD) or GRADSP (RAD) with or without caspase-1 (YVAD-CHO) and caspase-3 (DEVD-CHO) inhibitors for 4 h. Samples were

immunoblotted with anti-caspase-1, -2 or -3 antibodies. **c**, Cell extracts from Molt-4 cells were incubated with control beads or with GRGDNP- or GRADSP-coated beads. Samples were immunoblotted with anti-caspase-3 antibody. **d**, Pro-caspase-3 (M_r 32K) was immunoprecipitated and incubated without peptide (control) or with GRGDNP or GRADSP. Samples were immunoblotted with anti-caspase-3 antibody. **e**, Intact Molt-4 cells were treated for 18 h with 1 mM GRADSP or GRGDNP. **f**, Cell extracts from Molt-4 cells were incubated for various times in the presence of 1 mM GRADSP or GRGDNP or 10 μ M cytochrome c. **g**, 32K pro-caspase-3 isolated by immunoprecipitation was incubated with 1 mM GRADSP or GRGDNP for 1 h. **e-g**, Caspase-1 (open bar) and caspase-3 (black bar) enzymatic activity was then measured (mean $\pm$ s.d., $n = 3$).

peptides were able to induce DNA fragmentation and the characteristic morphological features associated with apoptosis (Fig. 2c). These results indicate that, in order to deliver their apoptotic signal, RGD peptides have to enter cells and are likely to have a non-integrin, intracellular target.

In an attempt to identify apoptosis-regulating proteins that might contain an RGD motif, we examined the Swissprot protein database for proteins containing a linear RGD sequence. We found an RGD motif next to the catalytically active cysteine residue in caspase-1 (QACR286GDSP, where 286 is the residue number of the amino acid preceding this number in this sequence). A homologous single RGD sequence is also found in caspase-2 (QACR304GD). In addition, the large subunit of caspase-3 and caspase-7 and the small subunit of caspase-8, also contain RGD sequence beginning at positions R144, R167 and R435, respectively (Fig. 3a).

The molecular requirements for activation of caspases are not well understood but are characterized by intricate activating and inhibitory protein-protein interactions¹³⁻¹⁵. The association of caspase precursors into oligomers before processing and activation indicates that factors controlling pro-enzyme aggregation may act as powerful regulators of apoptosis^{14,16}. Examples include the oligomerization of pro-caspase-8 by the adaptor molecule FADD during Fas-mediated apoptosis¹⁷, the forced crosslinking of caspases by chemically or proximity-induced oligomerization^{18,19} and the release of the APAF1-caspase9 complex from the anti-apoptotic protein Bcl-X_L by cytochrome *c* and dATP^{13,20}.

We therefore reasoned that RGD-containing peptides might act by inducing conformational changes in pro-caspases, leading to increased oligomerization and subsequent autoprocessing of these enzymes. This concept is supported by the presence of potential RGD-binding sites (D315DA and D180DM)⁴ near to the experimentally determined processing sites in pro-caspases 1 and 3, respectively (Fig. 3a). To examine this possibility we used cell extracts lacking plasma membranes. RGD peptides were able to induce caspase-3 activation as measured by the loss of caspase-3 proenzyme (relative molecular mass 32,000 (*M_r* 32K) and concomitant appearance of mature caspase-3 (*M_r* 17K) (Fig. 3b). This activation was blocked by a caspase-3 inhibitor (Asp-Glu-Val-Asp-aldehyde (DEVD-CHO)) but not by a caspase-1 inhibitor (Tyr-Val-Ala-Asp-aldehyde (YVAD-CHO)). Cleavage of caspase-1 and caspase-2 was not induced by RGD peptides (Fig. 3b). Intact nuclei added to this cell-free system showed chromatin condensation and typical apoptotic morphology following addition of RGD but not RAD peptides. Furthermore, RGD-coated beads were also able to induce caspase-3 activation in the cell-free system, showing that these coated beads, which were unable to induce apoptosis when exposed to intact cells (Fig. 2a), were functionally active when exposed to cytoplasmic extracts (Fig. 3c).

The results indicated that RGD-containing peptides might exert their effects by specifically and directly activating caspase-3. To test this hypothesis, we purified pro-caspase-3 from cytoplasmic extracts by immunoprecipitation and examined the ability of RGD-containing peptides to induce autoprocessing. When incubated in the presence of RGD-containing peptides pro-caspase-3 underwent rapid autoprocessing, leading to the appearance of the mature 17K subunit (Fig. 3d). To confirm that RGD peptides induce functionally active caspase-3, we measured caspase-3 and caspase-1 catalytic activity in intact cells, in cell lysates and in purified immunoprecipitated pro-caspase-3. In all cases, RGD but not RAD peptides were able to activate caspase-3 activity, but neither peptide induced caspase-1 activity (Fig. 3e-g). As the primary processing site between the p12 and p17 subunits in pro-caspase-3 lies next to the sequence DDM, a potential RGD-binding site³, we suggest that RGD peptides exert their effects by binding to this site, inducing conformational changes that lead to autoprocessing of caspase-3.

To confirm the requirement for caspase-3 in RGD-mediated cell

death, we used the breast carcinoma cell line (MCF-7) in which there is a functional deletion in the caspase-3 gene²¹. We stably transfected these cells with vector alone or with caspase-3 (Fig. 4a). RGD-containing peptides were unable to induce apoptosis in either parental cells or vector-alone transfectants. However, caspase-3-containing transfectants underwent rapid cell death when treated with RGD peptides (Fig. 4b). Exclusion of propidium iodide was used to detect apoptosis, because parental MCF-7 cells cannot cleave their DNA but can undergo caspase 3-independent programmed cell death²¹. These findings show that caspase-3 is required for RGD-mediated apoptosis.

Numerous studies have shown that the disruption of cell-matrix interactions leads to apoptosis in anchorage-dependent cells². The mechanism for this form of detachment-induced cell death, or 'anoikis', remains unclear, but has been ascribed to the loss of integrin-mediated survival signals². In many of these studies RGD-containing peptides induced rapid apoptosis^{22,23}. It has been assumed that this was because of the disruption of integrin-ligand interactions. An alternative explanation, given our observations in leukocytes, could be that these RGD peptides induced apoptosis through direct caspase-3 activation. To investigate this possibility, we studied the effect of RGD peptides on the survival of lung fibroblasts cultured on type 1 collagen, where cell attachment is RGD-independent²⁴. We observed cell death (Fig. 5a) and active caspase-3 (Fig. 5b) within 24 hours of treatment with RGD peptides. In addition, pro-caspase-3 cleavage was detected in both adherent and detached cells treated with RGD peptides (Fig. 5c). There was also a marked increase in the number of apoptotic adherent fibroblasts treated with RGD peptides compared with the number of apoptotic fibroblasts treated with RAD (Fig. 5d, e). These data indicate that RGD-containing peptides are able to induce apoptosis before these adherent cells detach from their substrate. Thus, as is the case for leukocytes, direct induction of

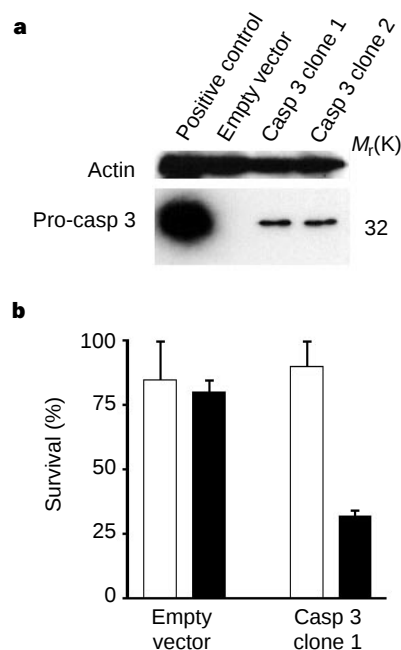


Figure 4 Caspase-3 is required for RGD-mediated cell death. **a**, Parental MCF-7 cells were stably transfected with caspase-3 clones 1 or 2 or pcDNA-3 vector alone, and caspase-3 expression was confirmed by immunoblotting. LCL cells were used as a caspase-3-positive control and β -actin as a gel-loading control. **b**, Vector-alone or caspase-3 transfectants were cultured in the presence of 1 mM GRADSP (open bar) or GRGDNP (black bar) for 18 h. Cell survival was measured by propidium iodide exclusion and flow cytometry (mean $\pm$ s.d., $n = 3$) and is expressed as percentage survival of input cells.

caspase-3 activation appears to be a major effect of treating these fibroblasts with RGD peptides, independent of any effects that these peptides may have on cell adhesion.

Peptides based on the RGD motif have been used extensively to inhibit tumour metastasis⁷, abrogate T-cell-mediated immune responses *in vivo*⁸, inhibit gastrulation in amphibian embryos²⁵ and induce endothelial-cell apoptosis in models of angiogenesis⁹. However, the exact mechanisms by which these peptides exert their biological effects *in vivo* have remained unclear, particularly as the observed therapeutic half-life of RGD peptides is longer, and their efficacy greater, than would be expected if their sole function was to inhibit integrin–ligand interactions^{1,7}. Our results therefore challenge the prevailing paradigm that RGD peptides work only through effects on integrin-mediated cell adhesion^{1,26}. We suggest that direct induction of apoptosis through caspase-3 activation is a major consequence of treating cells with these peptides.

The *in vivo* relevance of these peptides remains to be elucidated.

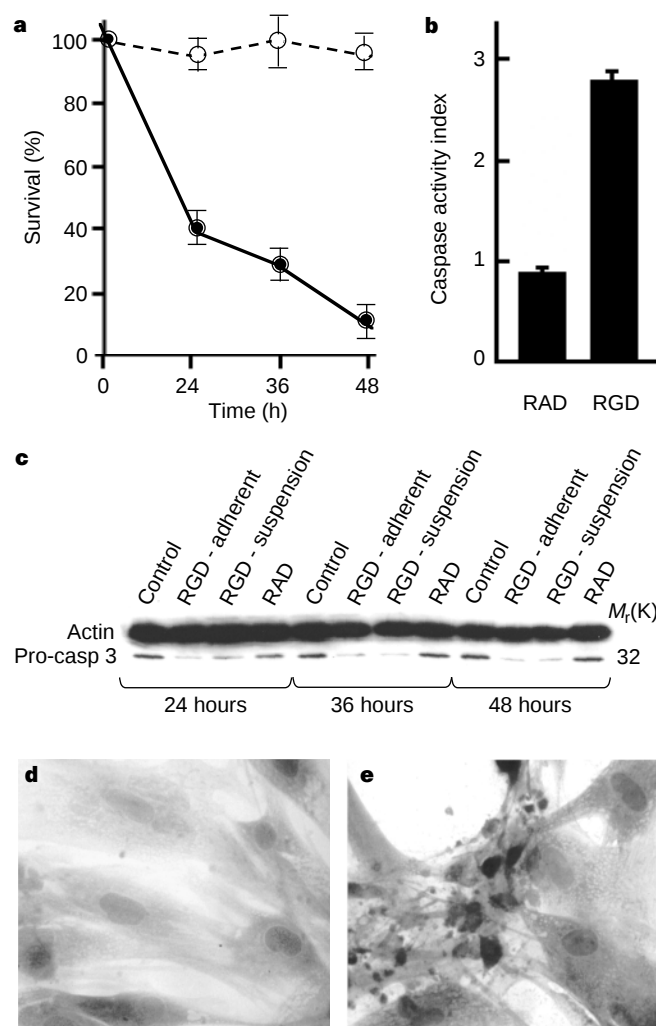


Figure 5 RGD peptides induce apoptosis and caspase-3 activation in lung fibroblasts cultured on type I collagen. Fibroblasts were cultured in the presence of 1 mM GRADSP (open circles) or GRGDNP (black circles) for various times. **a**, Cell survival, compared with untreated cells, was measured by trypan blue exclusion (mean $\pm$ s.d., $n = 3$). **b**, Caspase-3 activity was measured at 24 h. **c**, Caspase-3 cleavage was measured by immunoblotting. For RGD-treated cells, both adherent cells and suspension cells were collected separately. **d**, **e**, Fibroblasts were incubated with **d**, 1 mM GRADSP or **e**, 1 mM GRGDNP for 24 h before fixation and staining. Representative fields (original magnification $\times 40$) are shown.

However, several key regulatory proteins involved in the initiation of apoptosis contain potential RGD-binding sequences (DDX) at critical sites of protein–protein interaction. For example, APAF-1, the human homologue of the *Caenorhabditis elegans* death protein CED-4, contains six DDX sites²⁷. Five of these are within WD40 repeats thought to be involved in protein–protein interactions¹³. The sixth site is the conserved D250DV region in APAF-1/CED-4; this region is needed for the induction of CED-4 oligomerization and subsequent apoptosis^{19,27,28}. Furthermore, the Tat protein of HIV-1 contains an R78GD motif within its sequence; it is rapidly internalized by cells and can induce apoptosis in uninfected bystander lymphocytes²⁹. We are now investigating the mechanisms by which RGD-containing peptides enter cells and whether such peptides are produced during proteolysis and remodelling of the extracellular matrix³.

Our results define a new and specific way of activating the apoptosis cascade. We describe for the first time, to our knowledge, a peptide motif that is able to promote caspase-3 activation directly. This peptide motif, RGD, acts as a trigger sequence for integrin activation and adhesion⁶ and also appears to have a central function in the apoptotic programme, analogous to other well-known protein motifs involved in signalling networks such as the leucine zipper or zinc finger, which are also functional in many different proteins. Finally, our results have direct implications for ongoing clinical trials that make use of RGD-based compounds. As these peptides directly activate apoptosis, they are likely to be especially helpful in situations in which leukocyte accumulation and resistance to apoptosis contribute to disease pathology.

Methods

Cell culture, separation and transfection. Peripheral blood lymphocytes and resting peripheral blood CD4⁺ T cells were isolated and cultured as described¹⁰. B-cell lines LCL-1 and LCL-2 (transformed by Epstein–Barr virus, EBV) were produced by standard procedures. Leukaemic T-cell lines (Jurkat and Molt-4) and the erythroleukaemic cell line K562 were obtained from ATCC. Cells were cultured (at concentrations varying from 10^3 to 10^6 cells per ml) in RPMI medium supplemented with 2 mM glutamine, penicillin (100 U ml^{-1}), streptomycin ($100 \mu\text{g ml}^{-1}$), insulin and transferrin (both at $5 \mu\text{g ml}^{-1}$) and sodium selenite (5 ng ml^{-1}). Normal human serum and fetal calf serum were used at the stated concentrations. When serum-free medium was used, 1% bovine serum albumin was added in place of serum. We used 10 mM HEPES to control pH carefully in all cell cultures. Embryonic lung fibroblasts were cultured as described¹⁰. The breast carcinoma cell line MCF-7 was obtained from M. Wakelam and transfected with full-length Flag-tagged caspase-3 in pcDNA-3 (vector provided by G. Nunez)³⁰ or the expression vector alone ($2 \mu\text{g DNA}$), using electroporation (5×10^6 cells, 450 V, 125 μF). We selected cells for stable caspase-3 expression at 40 h post-transfection in medium containing $400 \mu\text{g ml}^{-1}$ G418 (Life Technologies).

RGD peptides and coated beads. Peptides were added to cell cultures at the stated doses. RGD, RGDS and RGEs peptides were obtained from Sigma. GRGDSP, GRGDNP, GRADSP and YIGSR were obtained from Calbiochem. All peptides were resuspended in PBS and stored in aliquots at -70°C . Amino-terminal-biotinylated peptides b-GRGDSP, b-GRADSP and b-GRGDNP were purchased from Alta BioSciences (Birmingham, UK). All peptides were purified by high-performance liquid chromatography and were $>95\%$ pure. Streptavidin-coated Dynabeads (M-280, 2.8 μm diameter, Dynal) were coated with biotinylated peptides according to the manufacturer's instructions. ProNectin F-coated plates were purchased from Protein Polymer Technologies (San Diego, CA, USA).

Immunofluorescence and confocal microscopy. Molt-4 cells were incubated in the presence or absence of biotinylated peptides (GRADSP or GRGDNP) at 1 mM for 18 h before peptides were detected with streptavidin–fluorescein isothiocyanate (FITC). Cells were either washed and fixed with Permeafix (Ortho) and then stained with streptavidin–FITC, to assess labelling of the outside and inside of the cells, or were alternatively washed, stained with streptavidin–FITC and subsequently fixed with Permeafix to assess labelling of the cell surface. Cell smears were made before counterstaining with propidium

iodide ($1 \mu\text{g ml}^{-1}$). Cells were photographed using a confocal microscope (Biorad MRC 500/600 hybrid).

Detection of apoptosis and caspase protease activity. Cell death was analysed by the exclusion of propidium iodide ($5 \mu\text{g ml}^{-1}$) and by forward- and side-scatter characteristics¹⁰. Results are expressed as the percentage of live cells remaining compared to the starting cell number. Lung fibroblasts were cultured on bacteriological plastic or glass coverslips coated with type I collagen (Sigma) ($10 \mu\text{g ml}^{-1}$ in PBS) for 1 h before the addition of RGD-containing peptides. Viable adherent cells were counted using trypan blue exclusion and samples were taken with which to assess caspase-3 processing and enzymatic activity. Cells grown on glass coverslips were fixed in acetone and stained with haemalum to confirm apoptosis. Enzymatic caspase-1 and caspase-3 activity was measured using the CaspACE fluorometric assay system (Promega) according to the manufacturer's protocol. Caspase activation is shown as the ratio between the caspase activity of the sample and that measured in control (non-treated) samples; for all samples, the background activity remaining after inhibition by either caspase-3 (DEVD-CHO) or caspase-1 (YVAD-CHO) inhibitors was subtracted.

Cytoplasmic extracts, immunoprecipitation SDS-PAGE and western blot analysis. Cytoplasmic extracts were prepared from cells in lysis buffer (10 mM PIPES, pH 7.4, 10 mM KCl, 2 mM MgCl_2 , 1 mM dithiothreitol (DTT), 5 mM EGTA, $25 \mu\text{g ml}^{-1}$ leupeptin, $5 \mu\text{g ml}^{-1}$ pepstatin, 40 mM β -glycerophosphate, 1 mM phenyl methylsulphonyl fluoride). Cells were homogenized with a Dounce homogenizer and lysis was confirmed by microscopy. Nuclei were collected by centrifugation through 30% sucrose (800g, 10 min at 4°C); mitochondria were collected at 10,000g for 20 min at 4°C ; and remaining cell membranes were removed by centrifugation at 100,000g for 45 min at 4°C . The supernatant from this final centrifugation was used as the cytosolic extract. YVAD-CHO and DEVD-CHO (Promega) were used at $10 \mu\text{M}$ and $20 \mu\text{M}$, respectively. For western blot studies, whole-cell lysates were prepared with RIPA buffer containing protease inhibitors as above. After adjustment for protein concentration (BCA assay, Pierce, Chester, UK), cell lysates were boiled in Laemmli buffer and resolved on 15% SDS-PAGE before western blot analysis with antibodies against caspases 1, 2 and 3 (Transduction Laboratories), Bcl2 and BclX (PharMingen), and β -actin (Sigma). Pro-caspase-3 was purified by immunoprecipitation from lysates of Molt-4 cells by lysis in 10 mM Tris, pH 7.4, 0.15 M NaCl, 1 mM EDTA, 1 mM EGTA, 0.2 mM sodium orthovanadate, 0.5% NP40 and 1% Triton-X100, with protein inhibitors as above. Cytoplasmic extracts were incubated with anti-caspase-3 antibody for 1 h at 4°C . After addition of sheep anti-mouse IgG (Binding Site, Birmingham, UK) and incubation for a further 1 h at 4°C , protein A agarose was added and the immunoprecipitated material was washed three times in lysis buffer. The beads were then resuspended in 1 ml caspase buffer (10 mM HEPES, pH 7.0, 50 mM NaCl, 20% glycerol, 40 mM β -glycerophosphate, 2 mM MgCl_2 , 5 mM DTT) and incubated with 1 mM peptides (GRGDNP or GRADSP) for 10 min at 30°C . At the end of the incubation an equal volume of Laemmli buffer was added, and the samples were boiled and resolved on 15% SDS-PAGE before detection of pro-caspase-3-cleaved fragments using the original antibody to caspase-3.

Received 14 October; accepted 14 December 1998.

1. Ruoslahti, E. RGD and other recognition sequences for integrins. *Annu. Rev. Cell Dev. Biol.* **12**, 697–715 (1996).
2. Meredith, J. E. & Schwartz, M. A. Integrins, adhesion and apoptosis. *Trends Cell Biol.* **7**, 146–150 (1997).
3. Werb, Z. ECM and cell surface proteolysis: regulating cellular ecology. *Cell* **91**, 439–442 (1997).
4. Pasqualini, R., Koivunen, E. & Ruoslahti, E. A peptide isolated from phage display libraries is a structural and functional mimic of an RGD-binding site on integrins. *J. Cell Biol.* **130**, 1189–1196 (1995).
5. Nicholson, D. W. *et al.* Identification and inhibition of the ICE/ced-3 protease necessary for mammalian apoptosis. *Nature* **376**, 37–43 (1995).
6. Du, X. P. *et al.* Ligands activate integrin $\alpha\text{IIb}\beta_3$ (platelet gpIIb-IIIa). *Cell* **65**, 409–416 (1991).
7. Humphries, M. J., Olden, K. & Yamada, K. M. A synthetic peptide from fibronectin inhibits experimental metastasis of murine melanoma-cells. *Science* **233**, 467–470 (1986).
8. Ferguson, T. A., Mizutani, H. & Kupper, T. S. Two integrin-binding peptides abrogate T cell-mediated immune-responses *in vivo*. *Proc. Natl Acad. Sci. USA* **88**, 8072–8076 (1991).
9. Brooks, P. C. *et al.* Integrin $\alpha\text{v}\beta_3$ antagonists promote tumor-regression by inducing apoptosis of angiogenic blood-vessels. *Cell* **79**, 1157–1164 (1994).
10. Salmon, M. *et al.* Inhibition of T cell apoptosis in the rheumatoid synovium. *J. Clin. Invest.* **99**, 439–446 (1997).
11. Lwebuga-Mukasa, J. S. A Mn^{2+} -enhanced, RGD-dependent adhesion technique for isolation of adult rat type II alveolar epithelial cells for immediate functional studies. *Am. J. Resp. Cell Mol. Biol.* **10**, 347–354 (1994).
12. Yamada, K. M. & Miyamoto, S. Integrin transmembrane signaling and cytoskeletal control. *Curr. Opin. Cell Biol.* **7**, 681–689 (1995).
13. Jacobson, M. D. Programmed cell death: a missing link is found. *Trends Cell Biol.* **7**, 467–469 (1997).
14. Salvesen, G. S. & Dixit, V. M. Caspases: intracellular signaling by proteolysis. *Cell* **91**, 443–446 (1997).
15. Thornberry, N. A. & Lazebnik, Y. Caspases: enemies within. *Science* **281**, 1312–1316 (1998).
16. Henkart, P. A. ICE family proteases—mediators of all apoptotic cell death? *Immunity* **4**, 195–201 (1996).
17. Martin, D. A., Siegel, R. M., Zheng, L. X. & Lenardo, M. J. Membrane oligomerization and cleavage activates the caspase-8 (FLICE/mach α 1) death signal. *J. Biol. Chem.* **273**, 4345–4349 (1998).
18. MacCorkle, R. A., Freeman, K. W. & Spencer, D. M. Synthetic activation of caspases: artificial death switches. *Proc. Natl Acad. Sci. USA* **95**, 3655–3660 (1998).
19. Yang, X. L., Chang, H. Y. & Baltimore, D. Essential role of ced-4 oligomerization in ced-3 activation and apoptosis. *Science* **281**, 1355–1357 (1998).
20. Pan, G. H., O'Rourke, K. & Dixit, V. M. Caspase-9, Bcl-xL, and APAF-1 form a ternary complex. *J. Biol. Chem.* **273**, 5841–5845 (1998).
21. Janicke, R. U., Sprengart, M. L., Wati, M. R. & Porter, A. G. Caspase-3 is required for DNA fragmentation and morphological changes associated with apoptosis. *J. Biol. Chem.* **273**, 9357–9360 (1998).
22. Frisch, S. M. & Francis, H. Disruption of epithelial cell-matrix interactions induces apoptosis. *J. Cell Biol.* **124**, 619–626 (1994).
23. McGill, G., Shimamura, A., Bates, R. C., Savage, R. E. & Fisher, D. E. Loss of matrix adhesion triggers rapid transformation-selective apoptosis in fibroblasts. *J. Cell Biol.* **138**, 901–911 (1997).
24. Hayman, E. G., Pierschbacher, M. D. & Ruoslahti, E. Detachment of cells from culture substrate by soluble fibronectin peptides. *J. Cell Biol.* **100**, 1948–1954 (1985).
25. Boucaut, J. C. *et al.* Biologically-active synthetic peptides as probes of embryonic-development—a competitive peptide inhibitor of fibronectin function inhibits gastrulation in amphibian embryos and neural crest cell-migration in avian embryos. *J. Cell Biol.* **99**, 1822–1830 (1984).
26. Souza, S. E., Ginsberg, M. H. & Plow, E. F. Arginyl-glycyl-aspartic acid (RGD)—a cell-adhesion motif. *Trends Biochem. Sci.* **16**, 246–250 (1991).
27. Zou, H., Henzel, W. J., Liu, X. S., Lutschg, A. & Wang, X. D. APAF-1, a human protein homologous to *C-elegans* ced-4, participates in cytochrome c-dependent activation of caspase-3. *Cell* **90**, 405–413 (1997).
28. Chinnaiyan, A. M., O'Rourke, K., Lane, B. R. & Dixit, V. M. Interaction of ced-4 with ced-3 and ced-9: a molecular framework for cell death. *Science* **275**, 1122–1126 (1997).
29. Li, C. J., Friedman, D. J., Wang, C. L., Meteliev, V. & Pardee, A. B. Induction of apoptosis in uninfected lymphocytes by HIV-1 Tat protein. *Science* **268**, 429–431 (1995).
30. Hu, Y. M., Benedict, M. A., Wu, D. Y., Inohara, N. & Nunez, G. Bcl-xL interacts with APAF-1 and inhibits APAF-1-dependent caspase-9 activation. *Proc. Natl Acad. Sci. USA* **95**, 4386–4391 (1998).

Acknowledgements. We thank N. Shamsadeen, M. Zwede and D. Hardie for technical assistance. This work was funded by grants from the Arthritis Research Campaign (to D.P., D.S.-T., M.S.), The Royal Society (to J.M.L.), Cancer Research Campaign (to N.V.H.), Medical Research Council (to G.P., K.T., A.N.A.) and Wellcome Trust (C.D.B., D.L.S.).

Correspondence and requests for materials should be addressed to M.S. (e-mail: M.Salmon@bham.ac.uk).

Proteins and peptides

Made-to-order proteins, a system to seek out RNA–protein interactions, new kit for enzyme kinetics and PCR-grade proteinase-K are among this selection of recent products.

pBridge

From Clontech Laboratories

www.clontech.co.uk

Three-hybrid analysis for more screening options

Clontech's new vector, pBridge, used with the company's Matchmaker GAL4 two-hybrid systems, forms a three-hybrid system. This can allow detection of protein–protein interactions dependent on the function of a third accessory protein. An accessory protein can stabilize a weak protein interaction, modify one of the interacting proteins (a kinase, for example), or 'bridge' two proteins that do not directly interact. The vector can also be used to screen for proteins that interfere with a known biointeraction.

Reader Service No. 100

Hybrid Hunter

From Invitrogen

www.invitrogen.com

Detect protein–RNA interactions in vivo

The RNA-Protein Hybrid Hunter from Invitrogen is specifically designed for the *in vivo* detection of RNA–protein interactions. Based on established two-hybrid technology, the system uses a RNA–hybrid intermediate to detect interactions between a known protein, the 'prey', and a known 'bait' RNA.

Reader Service No. 101

Targeting oncoproteins

From BioGenex

www.biogenex.com

Research antibodies to oncoprotein peptides

BioGenex offers a wide selection of monoclonal and polyclonal antibodies for the study of human oncoproteins by immunohistochemistry, western blotting and EIA. The antibodies are raised against a synthetic peptide derived from the respective oncoprotein. New antibodies include those to the c-fos, c-jun, and c-kit oncoproteins.

Reader Service No. 102

HRPStabilizer and APStabilizer

From Kirkegaard & Perry Laboratories

www.kpl.com

Stabilizers for horseradish peroxidase and alkaline phosphatase conjugates

KPL announces two new conjugate stabilizers for use with HRP and AP conjugates. Conjugates can be diluted in HRPStabilizer or APStabilizer to a concentration of 1 µg per ml and are stable for up to 6 months at 2–8°C. These stabilizers extend conjugate shelf life, are easy to use and reliable. AP or

HRP conjugates diluted in these stabilizers are claimed to retain greater functional stability than 50% glycerol or BSA.

Reader Service No. 103

MVD/24

From Charles Supper Company

www.charles-supper.com

A 24-chamber vapour diffusion plate for growing protein and crystals

Designed as an easy-to-use alternative to Linbro plates for manual and automated systems, the MVD/24 Crystal Growth Chamber features 24 individual chambers, each with a centre support post with a cup for holding the protein drop surrounded by a narrow moat where the reservoir solution is held. This format provides a large surface area for free diffusion during equilibration. Each cup and post isolates the protein drop and the chamber configuration prevents splashing when plates are transported.

Reader Service No. 104

QCB bioactive peptides

From Autogen Bioclear

www.autogen-bioclear.co.uk

A range of bioactive peptides

The QCB range of bioactive peptides comprises: β-amyloid peptides, neuropeptide Y and peptide YY-related products, protein kinase C peptides, SH2/SH3 domain and tyrosine kinase peptides, cAMP-dependent protein kinase peptides, neuropeptides, apoptosis peptides, leptin fragments, MAP kinase peptides, mastoparan peptides, CaM kinase substrates and inhibitors and miscellaneous kinase and phosphatase peptides.

Reader Service No. 105

L-8800 analyser

From Hitachi Instruments www.hil.hitachi.com

Amino acid analysis at speed

Hitachi's new amino acid analyser uses the company's proprietary analytical column,



Ready, go: the L-8800 analyser allows rapid flow.

packed with 3-mm particles to allow rapid flow. The reaction column is packed with beads to ensure homogenous flow of the mobile phase during the post-column reaction step. The L-8800 data system is a 32-bit application running on Windows NT 4.0. Up to 15 analysis modes are supported.

Reader Service No. 106

Tag antibodies

From Zymed Laboratories www.zymed.com

Epitope tagging for the purification, identification and analysis of proteins

Zymed's new monoclonal and polyclonal antibodies react specifically with the octapeptide epitope tag sequence DYKDDDDK, at the C- or N-termini of a protein. Epitope tags are short peptide sequences, easily recognized by readily available tag-specific antibodies. Zymed offers monoclonal mouse and polyclonal rabbit anti-octapeptide epitope tag antibodies, both suitable for western blotting, immunoprecipitation and ELISA.

Reader Service No. 107

FLAG

From Sigma

www.sigma.sial.com

Recombinant protein expression

Sigma's FLAG Protein Expression System is an established method for the expression, purification and detection of recombinant proteins fused to the FLAG peptide epitope. Three expression hosts are offered: *E. coli*, mammalian and yeast. Call 800-282-1298 (in the United States) for a brochure

Reader Service No. 108

UV-Star

From Lambda Physik www.lambdaphysik.com

A new peptide analysis application for a tuneable laser

Tyrosine, phenylalanine and tryptophan contain aromatic chromophores that absorb strongly at wavelengths below 310 nm, making it difficult to differentiate between these three amino acids by laser-induced fluorescence (LIF) analysis. Lambda Physik's answer to the problem is the UV-Star tuneable solid-state laser, for absorption and LIF analysis of DNA and amino acid peptide components. It carries out absorption and LIF studies at wavelengths between 282 and 310 nm. At these wavelengths the absorption of radiation by tyrosine and phenylalanine declines sharply relative to tryptophan, allowing LIF quantification of tryptophan.

Reader Service No. 109

FT-NIR spectroscopy

From Perkin-Elmer www.perkin-elmer.com

FT-NIR spectroscopy for moisture and protein determination

A new application note from Perkin-Elmer describes the use of FT-NIR spectroscopy to determine, with an estimated prediction error of less than 0.5%, the protein and moisture content in ground wheat raw materials, as used in the agricultural industry. The technique offers a simple alternative to traditional quantitative methods, which often take many hours to provide results.

Reader Service No. 110

Proteinase-K

From Boehringer Mannheim

www.biochem.boehringer-mannheim.com

DNA-free PCR-grade proteinase-K enzyme

Proteinase-K cleaves and inactivates many native proteins, especially nucleases, and is commonly used for the release of high molecular weight nucleic acids from various sample materials. When the resulting DNA is required as a template for PCR it needs to be of high quality, intact, free of nicks and have undergone repeated extraction or purification steps. PCR-Grade Proteinase-K has been introduced to allow preparation of PCR templates free of nonspecific DNA.

Reader Service No. 111

Glycosylated purified proteins

From ProteinGenesys

A 'gene-to-protein' service

ProteinGenesys, which is based in Augst in Switzerland, is taking orders for purified, native glycosylated proteins. The customer names the protein, and Protein Genesys undertakes to perform the molecular biology and purification processes. Optional hosts can be specified — for example a protein can be produced from fission yeast or

mammalian cells for comparative studies.

Reader Service No. 112

Enzyme kinetics

VP-ITC

From MicroCal

www.microcalorimetry.com

A development in ultrasensitive isothermal titration calorimetry (ITC)

MicroCal's VP-ITC System is a universal detector for the study of binding reactions and kinetics. The VP-ITC will provide a complete thermodynamic profile of any molecular interaction including binding constant, stoichiometry, enthalpy and entropy in less than an hour. A new feature is user-selectable response times to permit use in 'passive' or 'high-gain' mode. Passive mode is useful when examining very small signal changes in systems having slow transients, while the high-gain mode allows rapid turnaround of more routine work.

Reader Service No. 113

IASys optical biosensor

From Labsystems Affinity Sensors

www.affinity-sensors.com

Titration analysis for enzyme kinetic studies

The IASys optical biosensor from Affinity Sensors provides a direct method for the accurate determination of the equilibrium constant of enzyme catalysis from the



Kinetic art: the IASys optical biosensor.

equilibrium response data. Application note 7.1 describes quantitative characterization of an enzyme inhibitor interaction. The stirred cuvette configuration of IASys permits the stepwise titration of the inhibitor onto the immobilized enzyme, allowing a complete kinetic profile to be obtained. This method is particularly useful when the immobilized ligand cannot be regenerated satisfactorily.

Reader Service No. 114

Instrumentation

Electrophoresis

From Syngene

www.syngene.com

Gel tanks come in all sizes

Syngene has introduced a range of gel tanks and electrophoresis equipment to complement its gel documentation and analysis systems. Horizontal gel tanks can be supplied in five different chamber sizes from a maximum of 20 × 25 cm down to 6.5 × 8.5 cm. Vertical gel tanks are available in Mini, Midi and Maxi versions, for gel sizes of 10 × 10 cm, 20 × 20 cm and 20 × 35 cm.

Reader Service No. 115

Genophore

From Techne

www.techneuk.co.uk

Combined electrophoresis

The Genophore system includes all the basic gel electrophoresis functions in a minimum of space. Gel preparation and subsequent electrophoresis are performed using an integrated power supply and gel tank. The cell separates simply from the power supply to permit cleaning of the gel tank. The tank is made of transparent meta-acrylic with a tray for gels 230 × 140mm and fluorescent rule.

Reader Service No. 116

These notes are compiled in the Nature office from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.

ADVERTISEMENTS

NEWCASTLE UNIVERSITY
MOLECULAR BIOLOGY UNIT



PROTEIN SEQUENCING

Membrane-bound, solid or solution samples

£12 per cycle

5 cycles minimum - no set up charge

For more information please contact Dr. Joe Gray:
Phone/Fax: +44 191 222 8612

Email: JOE.GRAY@newcastle.ac.uk

*DNA sequencing (ABI 377-96KL, 877 Catalyst)
also available at competitive prices*

Check out
our homepage at
<http://www.resgen.com>

Peptides

\$14/residue
10 residue minimum

Polyclonal Antibodies - \$1,195

Please contact us about complete
services for your custom peptide
and polyclonal antibody needs.

Research Genetics, Inc.
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 256-536-9016

Detect residual DNA with Wako's

DNA Extractor Kit

or Isolate Genomic DNA with

DNA Extractor WB Kit

**Both use the Nal Method
No Phenol No Chloroform**



Wako BioProducts
Wako Chemicals USA, Inc.
1600 Bellwood Road
Richmond, VA 23237
Telephone: (804) 271-7677
Facsimile: (804) 271-7791
(800) 992-9256
bioproducts@wakousa.com

READER ENQUIRY NO. 2

READER ENQUIRY NO. 4

READER ENQUIRY NO. 39

Expanding Canada's knowledge base

Canada needs scientists from almost all disciplines to fuel its economy. In the past, the country's attractiveness to job-seekers has been overshadowed by that of the United States, but things are changing.

Potter Wickware

Science is playing a key role in Canada's transition from an economy dependent on commodities and exports to one based on knowledge and technology. During this momentous change, the country is experiencing shortages — sometimes acute — of scientists in fields ranging from computing professions to information technology to health and biotechnology.

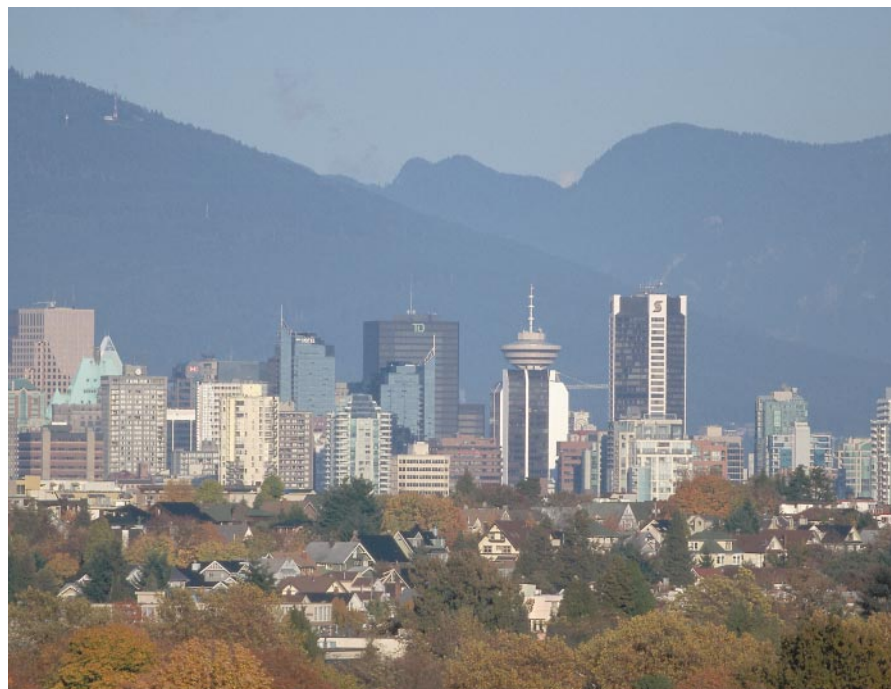
Federal and provincial governments and private industry are making concerted efforts to overcome present challenges and capitalize on potential as Canada progresses along this path.

"The science blanket is thin here," says Christoph Sensen, group leader of the Genomics Technologies Group at the National Research Council's Institute for Marine Biosciences in Halifax, Nova Scotia. "But the atmosphere is good and the community is thriving — and what scientists in Canada lack in resources, they make up in their ability to network and collaborate."

The jobs outlook in Canada is bright at present and looks even rosier in the future, according to Lise Hughes, at the Steacie Institute of Molecular Sciences in Ottawa, Ontario. "Most areas of the IT sector are in short supply, and recruitment efforts include those of companies like Nortel working closely with two universities, Carleton and Ottawa University," she says.

Academic jobs are also opening up as older professors retire, says Hughes. "Numerous openings for younger professors are being created thereby, for example at the University of Alberta, which has had nine openings this year alone."

David Cohen, a Montreal immigration lawyer, reports that — while Canada is currently experiencing a shortage of skilled workers in many areas of expertise —



Vancouver: closer to California's Silicon Valley by air (and in climate) than to much of Canada.

"computer professions are leaps and bounds ahead of the other occupations in demand".

Christoph Sensen estimates that, in Canada's diminutive but fast-growing field of bioinformatics, "there are at least 50 high-quality jobs with salaries of at least \$60,000 that are currently unfilled, and the demand is increasing very rapidly".

High-quality research

Canadian science compares well in quality with other advanced countries. As measured by citation impact, weighted according to a method described by J. Adams (*Nature* 396, 615–618; 1998), Canadian science ranks third in the world, behind the United States and England.

Although the number of published scientific papers from Canada is relatively low, their overall impact is high and is maintained across a broad front, compared with the other six countries that have major research and development (R&D) establishments (United States, England, France, Germany, Australia, Japan). Canada's top scientific sectors are shown in Table 1.

To put Canadian research jobs in perspective, though, it is useful to consider that Canada has only 8 per cent as many as the United States, which is perhaps not surprising given that — at 29 million — Canada's population is somewhat less than that of California. Total R&D personnel in Canada in

1995 came to 135,000, of whom 82,000 were researchers, more than 80 per cent of these concentrated in the areas of natural sciences and engineering.

In Canada, 54 per cent of researchers work in business and industry, 32 per cent in universities, 13 per cent in federal and provincial government and 1 per cent in non-profit organizations. With a labour force of 15 million, Canada has 55 researchers per 10,000 workers, somewhat below the 74 per 10,000 in the United States and well below Japan's 99.

Although concerted efforts are under way to increase the Canadian figures, a scarcity of both funding and people is impeding the build-up of the country's science establishment. With a lower per capita spending on science than most of other G7 countries, aggravated in recent years by cuts in federal payments to the provinces, government support to universities has decreased. Fees are accordingly going up, but even so the number of foreign students in Canada has fallen by 15 per cent to 32,000 in four years.

"In astronomy, there is a shortage of high-quality students," says Michael De Robertis, of the Department of Physics and Astronomy at Toronto's York University. "We imagine this to be partially the result of a depressed job market over the past few years." Undoubtedly it is also due to competition from Canada's huge and sometimes

Table 1 **Core strengths of Canadian science relative to world citation index (baseline= 1.0)**

Chemical engineering	1.29
Food science and technology	1.28
Clinical dentistry	1.24
Library/information management	1.23
Chemistry	1.19
Communications and media studies	1.18
Physics	1.18
Social policy & social work	1.16
Built environment	
(civil engineering and architecture)	1.13
Applied mathematics	1.13
Environmental sciences	1.12
Computer science	1.11

overbearing neighbour to the south, for he continues, "We have a serious postdoc shortage, chiefly the result of competition with American institutions — our salaries tend to be significantly smaller than theirs."

Spread of opportunities

Science and technology jobs are concentrated in six principal locations. Greater Montreal supports the telecommunications, aerospace, pharmaceutical and software industries. It has 575 companies which manage R&D projects totalling more than \$1 billion annually, with \$200 million of it in telecommunications. Computers, electronics and telecommunications account for \$2.8 billion of combined output and 21,500 of the region's jobs; the area also supports 350 software firms. US and European biotech companies have established Canadian headquarters in Montreal, including Astra, Phoenix International, Biochem Pharma, Algène and the Netherlands company Bio-Intermediar.

Most of Canada's foreign-owned companies have their headquarters in Greater Toronto. The city has three universities as well as research centres for companies including IBM and Xerox. In addition to telecommunications, approximately 40 per cent of the pharmaceutical companies in Canada are in Toronto, including the vaccine maker Pasteur Mérieux Connaught, Eli Lilly, Glaxo-Wellcome, Allelix, the bioinformatics company Base4 and others. The University of Toronto, having a large medical school with 32 medical research institutes, is a major centre of medical research.

In Guelph-Waterloo-Cambridge, Canada's 'technology triangle', Waterloo University is strong in software R&D, performing more contract research than any other university in Canada, according to a report by the auditing firm DeLoitte & Touche. Gloucester, Ontario, is home to Amita, a software company which has developed COSMOS, a networked database for automating Canada's consular services at more than 100 locations worldwide. Gloucester has about 65 high-tech firms, including Telesat Canada, Vitana Corporation and Hexagon Computer Systems.

Ottawa-Hull, Canada's 'Silicon Valley north' — centred on telecommunications and satellite technology — is home to Nortel Networks, Corel, Cognos, Newbridge Networks, Mitel, Digital Equipment and nearly 700 other technology companies, which in 1997 generated \$8 billion for the region.

In the prairie provinces, Calgary, Alberta, is home to Nortel's world centre for wireless technology, advanced technology related to oil exploration, distance education and multi-media. High-tech employment in Calgary is around 29,000. Edmonton, Alberta, is home to the biotech company Biomira, which specializes in cancer vaccine research. In Saskatoon, Saskatchewan, more than 200

Table 2 Occupational groups in key growth sectors

Information systems and data processing managers
Managers in health care
Physicists and astronomers
Chemists
Geologists, geochemists, geophysicists
Meteorologists
Other professional occupations in physical sciences
Biologists and related scientists
Forestry and agriculture professionals
Engineers (civil, mechanical, electrical and electronic, chemical, industrial manufacturing, metallurgical, materials, mining, geological, petroleum, aerospace)
Mathematicians and statisticians
Computer systems analysts and programmers
Health professionals

staff and guest researchers at the NRC's Plant Biotechnology Institute carry out research in plant genomics, gene expression and promoter technology, which has resulted in improved varieties of wheat, barley, flax, canola and other crops.

The biotechnology industry is thriving everywhere in Canada, but especially in British Columbia, where, according to a survey by the international auditing firm PricewaterhouseCoopers last March, it is growing by 20–30 per cent each year.

Home to about 1,000 of Canada's 16,000 technology companies, Vancouver, in British Columbia, is strong in communications, electronics, biotechnology, computers, alternative fuels, software and databases for hospital and finance. Vancouver enjoys a moderate climate and a time zone shared with California and the state of Washington, where many of its technology companies have branch offices. Only three and a half hours away by air, Vancouver is substantially closer to San Diego, and closer still to Silicon Valley, than to Montreal, which lies four time zones and nearly 4,000 miles to the east.

Attracting new talent

The Government is eager to see growth, particularly in information technology, telecommunications, electronics and computing, health and biotechnology, agriculture and food, and environmental and ocean technology. In microbiological research, for example, public grants to the Canadian Bacterial Disease Network (a consortium of researchers and developers from Government, academia and industry), augmented by support from private industry, have resulted in the creation of ten new companies, two of which are now publicly listed: Micrologix, of Vancouver, which focuses on the increasing problem of antibiotic resistance; and Synsorb, of Calgary, which produces carbohydrate-based neutralizers of bacterial toxins.

Canada is also endeavouring to attract the skilled talent it needs to staff new technology ventures and enlarge existing ones. In

Table 3 URLs of useful web sites

Canada Employment links, by Campbell-Cohen
http://canadavisa.com/dbank/employe.htm
Project to attract highly skilled temporary workers
http://cicnet.ci.gc.ca/english/press/98/9853-pre.html
NRC careers page
http://hr.nrc.ca:8080/HRB/careersm.nsf/pagee/home
NRC current jobs
http://hr.nrc.ca:8080/hrb/jobpost.nsf/poste
BC Biotech Alliance Employment Opportunities
http://www.biotech.bc.ca/bcaba/fset_about.html
NRC - Government labs
http://www.nserc.ca/programs/secl.htm
Biotechnology Human Resources Council National Job Bank
http://www.biotech.ca/bhrc/bhrcjob.html
Canada Technology Network (CTN)
http://ctn.nrc.ca/ctn/ctn.html
CTN members
http://data.ctn.nrc.ca/nat/navigate/byname/list-n-e.htm
Plant Biotechnology Institute
http://www.pbi.nrc.ca/
Public Service Canada IT recruitment
http://www.psc-cfp.gc.ca/it-ti/index.html

January, for example, Canada's Ministry of Citizenship and Immigration proposed an accelerated list-based method for granting work permits to foreign nationals, based on labour market information, rather than the present case-by-case method.

"The speed with which the modern economies move creates an urgency regarding entry of key personnel," said Minister Lucienne Robillard, taking note of the fact that employers around the world are competing for highly trained individuals to staff knowledge-based industries.

To make entry easier for skilled foreign workers, a pilot project was launched last October to grant work permits automatically to spouses accompanying temporary workers coming to Canada for jobs in key high-growth sectors of the economy (Table 2). Of the approximately 170,000 foreign workers who were issued employment authorizations by the Ministry of Citizenship and Immigration in 1997, 10,000 would have qualified under this pilot scheme.

Potter Wickware is a science writer in Oakland, California, USA.

e-mail: wickware@worldnet.att.net

Careers and recruitment in Nature

The Careers and Recruitment features published in *Nature* are collected on our web site (www.nature.com). They include:

Geosciences	3 Dec 1998
Academic-corporate partnerships	24 Sep 1998
Far East emergent economies	6 Aug 1998
Alternative careers for scientists	4 Jun 1998
Cancer research	26 Mar 1998
Cell biology	19 Feb 1998

Future features will include opportunities in physics, chemistry and clinical research, together with a survey of senior appointments and of science in the Benelux countries. Suggestions for topics to be covered within these subjects can be sent to Maxine Clarke at m.clarke@nature.com.